【保险精算】

考虑老年痴呆症的医疗险住院费用预测与比较

——基于机器学习模型

赵颖旭　包竹青　高　珊　周　亮　刘逸圣　陈　浩　张　宁

【摘　　要】老年痴呆症严重影响老年人生活质量,其影响程度也随着中国老年人口比例提升和预期寿命增加而不断加深。在此背景下,针对老年痴呆症的商业健康险开发意义重大。本文基于详实的历史数据,利用传统定价中估计定价因素的广义线性模型对老年痴呆症患者住院费用进行估算和预测,同时用随机森林、LASSO 回归、LightGBM 三种机器学习方法进行同样的估算预测,以期为保险公司开发产品、丰富保险市场、助力养老提供基础。对四种方法进行横向比较的结果显示,机器学习方法除了在结果估算方面具有更大的优势外,还在影响因素的数量、精准定价以及模型适应方面具有更大的潜力。最终的计算结果和模型使用的过程方法结合住院率数据可用于具体测算老年痴呆症医疗保险费用。

【关 键 词】老年痴呆症;商业医疗险;带病投保;机器学习;保费测算

【作者简介】赵颖旭,泰康健康产业投资控股有限公司首席医疗分析师,流行病学博士,中央财经大学中国金融科技研究中心兼职研究员;包竹青,泰康保险集团股份有限公司运营企划主管;高珊,泰康保险集团股份有限公司运营企划专员;周亮,中央财经大学中国金融科技研究中心兼职研究员;刘逸圣,泰康健康产业投资控股有限公司实习生;陈浩,中央财经大学保险学院研究生;张宁(通讯作者),中央财经大学金融学院教授,中央财经大学中国金融科技研究中心主任,E-mail:zhang-ning@vip.163.com。

【原文出处】《保险研究》(京),2020.9.64~76

【基金项目】本文得到中央高校基本科研业务费专项基金和教育部人文社会科学重点研究基地重大项目(16JJD790060)资助。

一、引言

老年性痴呆是老年人群的常见疾病,阿尔茨海默病是老年性痴呆最常见的形式,占痴呆症总病例的60%~70%(Organization,2019)。有研究显示,2010 年中国痴呆症患者数已达到 961 万人,预期将持续增长,2020 年预计达到 1406.9 万人,2030 年将达到 2329.1 万人(Wu,et al.,2017)。目前中国 60 岁及以上老年人群痴呆症患病率约为 4.03%,其中阿尔茨海默病患病人数占比超过 60%(Zhu,et al.,2019)。该疾病病程较长,合并症发病风险增加(Moon,et al.,2019),医疗服务利用及医疗费用显著增加(Pyenson,et al.,2019),给社会经济带来沉重负担。中国 2010 年痴呆症疾病经济负担已达到 472 亿美元,预计 2020 年将达到 690 亿美元,2030 年达到 1142 亿美元(Xu,2017)。医疗保险对老年人在医疗方面的总支出有显著促进作用。老年人的医疗支出在医疗保险的促进下增加了 41.5%,增加值为 1227 元。医疗保险显著提高了老年人的健康水平和及时就医的概率(梁志胜,2017)。

中国阿尔茨海默病患者的直接医疗费用占其

总疾病经济负担的32.51%,且显著高于美、英、法、德等发达国家及世界平均水平(Jia,et al.,2018)。中国痴呆症患者主要以居家或社区照护形式为主(Wang,et al.,2018),对痴呆症患者的长期照护为照护人带来巨大的精神压力、经济和心理负担(Kelley,et al.,2015;韩颖,2017;韩颖等,2016;康昊昱,2010;雷婷,2011)。有研究表明,基本医疗保险对于降低老年人自付医疗支出具有显著影响,但商业保险的覆盖仍然不足(梁志胜,2017)。

2007年8月1日,严重阿尔茨海默病被纳入《重大疾病保险的疾病定义及使用规范》(中央政府门户网站,2007)。大陆重疾险基本都已包含该病种,但部分产品限定确诊时间必须早于70周岁,自主生活能力完全丧失,无法完成六项基本日常生活活动中的三项或三项以上才满足赔付条件,对被保险人实行定额赔付(徐贝尔,2016)。

作为基本医疗保险的重要补充,商业医疗险拥有价值杠杆和风险防范的双重功能。近年来,百万医疗险逐渐风靡市场,以其低保费、高保额、高杠杆率吸引了消费者(董斌,2018;孙东雅、张铭哲,2019;王硕,2018)。但由于基于医疗费用实际发生率的精算定价基础相对缺乏,保费制定无差异化,同时通过互联网渠道进行产品投放时,目标偏向年轻标准体承保,通常不保证续保,使得带病投保和年龄限制成为产销阻力(潘兴,2014)。在商业医疗险加速发展的形势下,现阶段针对老年性痴呆的商业医疗险仍处于空白。

针对老年性痴呆进行医疗险差异化保费测算,能够在老龄化背景下突破年龄限制挖掘需求,并有效分散潜在风险,积极引导就医行为(舒晓燕,2011),规范治疗干预手段,降低疾病经济负担。目前,机器学习预测模型已被应用于风险预测、保费测算等产品设计环节中(Huang,et al.,2020;曾宇哲等,2019;韩耀风等,2017;郝君,2018;贾延延、冯键,2020;李红梅等,2020;李阳等,2020;林鹏程、唐辉,2019;孟生旺、黄一凡,2018;夏涛等,2019;张碧怡等,2019;张亦鼎,2019;孟生旺,2012;孟生旺等,2017),而应用机器学方法估算医疗费用,也已经有不少实践基础(冯菁楠等,2019;王文文,2016;夏涛,2019)。本研究对老年痴呆医疗险的差异化保费测算进行初步探索,以期对不同带病投保人的差异化保费定价有所助益。

本文第二部分主要介绍机器学习方法在损失厘定时的作用;第三部分介绍数据来源和模型方法;第四部分展示不同模型的拟合结果并对其进行对比;第五部分根据得到的结果给出结论。

二、研究方法介绍

(一)保费测算

在住院率与索赔强度相互独立的假设下,将索赔频率的预测值与索赔强度的预测值相乘得到纯保费的预测值。在这种建模方式下,每份保单的累积赔款可以表示为 $Y = X_1 + X_2 + \cdots X_N$,其中N表示索赔次数,$X_i(i = 1,2,\cdots,N)$表示第i次索赔的赔款金额,通常假设每次的赔款金额独立同分布,且与索赔次数N相互独立。

在这种假设下,纯保费可以表示为:

$$E(Y) = E(N)E(X) \quad (1)$$

在保险定价中,损失厘定常用的方法是广义线性模型(Generalized Linear Models,GLM)。GLM模型在20世纪70年代由Nelder和Macullagh引入精算学(Nelder and Wedderburn,1972),90年代由英国精算师首先应用于非寿险定价,并成为主要手段,例如在车险定价中用于风险预测(许译芝,2019)。

(二)GLM模型

广义线性模型是损失预测的主流方法(许译芝,2019;张碧怡,2019),是在一定的分布假设下建立的预测模型,要求损失数据满足一定的假设条件,譬如索赔次数服从泊松分布或负二项分布,索赔强度服从伽马分布或逆高斯分布,累积赔款服从Tweedie分布。GLM采用python3.8.3软件的GLM模块进行拟合,未加入交互因素项。

GLM的一般表达式如下,其中$g(\mu)$为连接函数。

$$h(x) = E[y|x,w] = \mu = g - 1(\eta) \quad (2)$$

当哑变量太多时,GLM的拟合结果可能效率不高。比如,在带病投保住院费用的测算过程中,很

难将上百种不同疾病组作为参数纳入模型，而不同疾病组显然对医疗费用有较大影响。所以考虑引入机器学习算法来解决传统算法不能解决的问题。常用的机器学习方法包括随机森林、LASSO 回归、LightGBM 三种，在 Windows 系统环境下使用 python3.8.3 软件。

（三）随机森林

随机森林由 Leo Breiman 和 Adele Cutler 提出，是利用多个决策树对样本进行训练、分类并预测的一种算法。在随机森林中，每一个决策树“种植”和“生长”的规则如下所示：

1. 假设设定训练集中的样本个数为 N，然后通过有重置的重复多次抽样来获得这 N 个样本，此抽样结果将作为生成决策树的训练集；

2. 如果有 M 个输入变量，每个节点都将随机选择 m（m < M）个特定的变量，运用这 m 个变量来确定最佳的分裂点。在决策树的生成过程中，m 保持不变；

3. 每棵决策树都最大可能地进行生长而不进行剪枝；

4. 通过使用 argmax（Var - VarLeft - VarRight）作为评判标准，即使得当前节点训练集的方差 Var 减去左子节点的方差 VarLeft 以及右子节点的方差 VarRight 值最大。

随机森林能处理高维特征，不容易产生过拟合，模型训练速度比较快，特别是对于大数据而言尤其如此，且对数据集的适应能力强：既能处理离散型数据，也能处理连续型数据，数据集无需规范化。该算法可以对数据进行分类并给出影响因素的排序，评估各个变量在分类中所起的作用，已经被广泛应用于保险领域，如保险购买预测场景，风险因子重要性测度，非寿险准备金相关测算等（郝君，2018；林鹏程、唐辉，2019；张碧怡，2019；安磊等，2016）。

（四）LASSO 回归

LASSO（Least Absolute Shrinkage and Selection Operator，LASSO）（Tibshirani，2011）方法于 1997 年由 Tibshirani 提出，是以缩小变量集（降阶）为思想的压缩估计方法。它客观筛选有效变量，构造一个惩罚函数，让回归系数绝对值之和在小于一个常数的约束条件下进行优化，最终使得回归模型残差平方和最小，从而有效解决回归模型中的多重共线性问题，进而达到变量选择的目的。

Lasso 回归是在损失函数后，加 L1 正则化，使得下式取最小值：

$$\frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\lambda\sum_{j=1}^{\kappa}|\omega_j|\right] \tag{3}$$

m 为样本个数，κ 为参数个数，其中 $\lambda\sum_{j=1}^{\kappa}|\omega_j|$ 为 L1 正则化。

Lasso 的复杂程度由 λ 来控制，λ 越大则对变量较多的线性模型的惩罚力度就越大，最终能进入模型的变量就越少，目前已广泛应用在医药卫生领域，用于住院费用的估算（Huang，2020；韩耀风，2017；李阳，2020）。

（五）LightGBM

LightGBM 算法是一种基于 GBDT（Gradient Boosting Decision Tree，GBDT）的数据模型，算法中使用回归树作为弱学习器，通过将每个预测结果与目标值的残差作为下一个学习的目标，获得当前残差回归树，每个树都学习所有先前树的结论与残差，将多个决策树的结果加在一起作为最终预测输出。

利用梯度单边采样的直方图算法对特征进行预排序，将样本某一特征上的单梯度作为样本的权值进行训练，并利用节点展开方式进行树的构建，是一种高效、高精度、高性能的分类算法。该算法的实现步骤如下：

1. 训练样本个数为 N，选取前 a% 个较大梯度的值作为大梯度值的训练样本；

2. 从剩余的 1 - a% 个较小梯度的值中，随机选取其中的 b% 个作为小梯度值的训练样本；

3. 对于较小梯度的样本，也就是 b% * N，在计算信息增益时将其放大（1 - a）/b 倍。

总的来说就是 a% * N + b% * N 个样本作为训练样本。而这样的构造是为了尽可能保持与总的数据分布一致，并且保证小梯度值的样本得到

训练。

LightGBM 具有较优的数据分类能力,且对于大量训练样本不容易过拟合。当前已应用于租金预测(陈熙、张晓博,2020),信用风险评级(马晓君等,2018),医学疾病预测(王悦等,2019;吴绍武、续育茹,2019)。

三、数据与模型构建

(一)研究对象

本研究对老年痴呆患者的定义为:主要诊断或其他诊断 ICD-10 编码为表 1 的住院患者(牛犇,2017;王莹等,2006)。国际疾病分类(International Classification of Diseases,ICD)是一种对不同类型疾病及健康相关问题进行编码和分类的国际标准,自产生至今已有上百年的历史。根据世界卫生组织和我国卫生健康委员会的要求,我国自 1987 年起推广应用 ICD-9(ICD 第 9 次修订本),从 2002 年起改为使用 ICD-10(ICD 第 10 次修订本),并一直沿用至今(贾友波、宋宪锟,2020;吕国友等,2019)。

本研究选取 2015~2017 年来自全国 30 个省、自治区和直辖市的数据,包括来自综合医院、脑科专科医院、精神专科医院、其他专科医院等超过 600 家医院的 101341 住院人次。数据中住院人次涉及的不同医院类别分布情况见表 2。

患者住院费用的数据年份分布见表 3 所示。可以看到,2015 年以来住院人次数持续增长,2015~2017 年各年度住院人次数分别占总住院人次的 18.73%、31.71% 和 47.3%。研究对象性别分布如表 4 所示,其中男性占 59.58%,共计 60379 人次;女性占比 36.44%,共计 36926 人次。图 1 给出了研究对象的年龄分布,可以看到,研究对象的平均年龄为 80.37 岁,中位数年龄为 82 岁,其中最年长的患者 99 岁。表 5 给出了患者地区分布统计情况:患者就医医院地区中,华东地区患者占比最大,共 31411 人次,占比 31.00%,东北患者数最少,共计 3546 人次,占比 3.50%。从上述统计描述可以看出,本研究样本量较大,且在时间、空间上分布均衡,具有代表性。

患者的疾病诊断 ICD-10 编码情况分布情况如表 6 所示:若患者主要诊断和其他诊断存在阿尔茨海默痴呆和血管性痴呆,则此患者以其主要诊断为准计入;若患者主要诊断非痴呆相关,而其他诊断同时存在阿尔茨海默痴呆和血管痴呆两类,则按照诊断顺序靠前的类别计入;最终阿尔茨海默痴呆相关的 ICD-10 诊断患者占比 61.61%,共计人次 62436 人,血管性痴呆患者占比 38.39%,共计 38905 人。

表 1　　研究对象 ICD-10 编码范围统计表

疾病分类	主要编码	附加编码	疾病名称
血管性痴呆	F01.0		急性发作的血管性痴呆
	F01.1		多发脑梗死性痴呆
	F01.2		皮层下血管性痴呆
	F01.3		混合型皮层和皮层下血管性痴呆
	F01.8		血管性痴呆,其他的
	F01.9		未特指的血管性痴呆
阿尔茨海默痴呆	F03.x		未特指的痴呆
	G30.8	F00.2*	其他阿尔茨海默病性痴呆
	G30.9	F00.9*	阿尔茨海默病性痴呆
		F00.0*	早发性阿尔茨海默病性痴呆
		F00.1*	晚发性阿尔茨海默病性痴呆
		F00.2*	阿尔茨海默病性痴呆,非典型或混合型

表 2 研究对象来源分布统计表

医院类别	住院人次	人次占比	医院数	医院数占比
综合医院	92188	90.97%	513	85.50%
脑科/精神科专科医院	926	0.91%	4	0.67%
其他专科医院	8227	8.12%	83	13.83%
合计	101341	100.00%	600	100.00%

表 3 住院费用数据年份分布统计表

年份	人次数	占比
2015	18985	18.73%
2016	32132	31.71%
2017	47938	47.30%
其他	2286	2.26%
合计	101341	100.00%

表 4 研究对象性别分布统计表

性别	人数	占比
男	60379	59.58%
女	36926	36.44%
不详	4036	3.98%
合计	101341	100.00%

表 5 患者地区分布统计表

地区	病例数	占比	地区	病例数	占比
华东	31411	31.00%	西南	18513	18.27%
华北	15057	14.86%	西北	4918	4.85%
华南	12282	12.12%	不详	626	0.62%
华中	14988	14.79%	合计	101341	100.00%
东北	3546	3.50%			

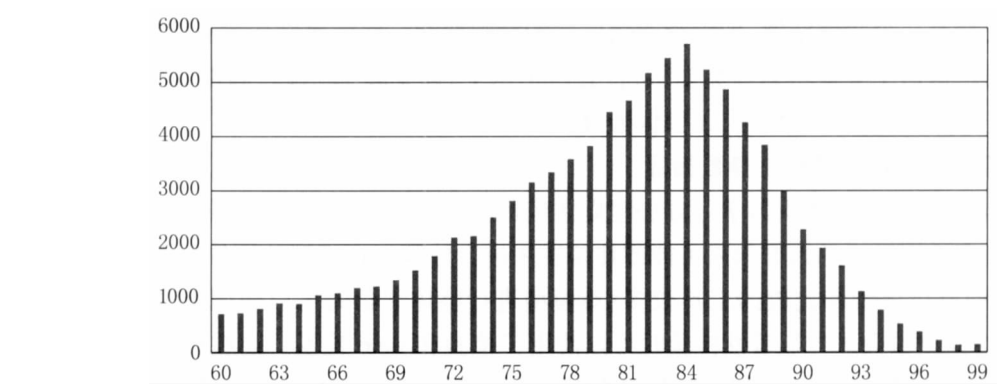

图 1 研究对象年龄分布图

表 6 患者疾病诊断分布情况统计表

疾病类型	主要诊断	其他诊断	合计	占比
阿尔茨海默痴呆	15373	47063	62436	61.61%
血管性痴呆	8243	30662	38905	38.39%
合计	23616	77725	101341	100.00%

（二）住院费用测算指标

拟合住院费用模型的预测指标选取三类，分别是保单属性、疾病属性和客户属性，各项指标的变量类型及水平见表 7 所示。

保单属性变量有一个为固定值，即预测费用发生的时间期间。疾病属性变量有两个，分别是诊断类型和疾病种类。这两个变量均为分类变量，各有两个水平，详见表 7。客户属性变量有四个，其中被保人年龄为连续变量，省份、性别、有无社保为分类变量。

（三）模型应用过程描述及评价参数指标

本文分别利用 GLM 方法和机器学习方法建立住院费用预测模型，模型通过保单、疾病、客户三个维度的变量对具有不同特征的人群未来可能的住院费用进行估算。在模型的评价方面，GLM 模型应用较成熟，对模型进行评价和比较的工具较多：对模型进行整体评价时可以使用 R^2、Deviance、AIC、BIC 和 F 检验，对模型参数进行显著性检验时可以使用 T 检验和 Wald 检验。而在机器学习的模型评价中，由于机器学习方法原理差异较大，模型理论假设比较少，因此对模型进行评价的方法较少，一般是通过比较模型预测结果与真实结果来判断模型的预测精度。

四、实践结果与对比

（一）GLM 模型结果及分析

GLM 模型中考虑的变量包括研究对象的年龄、省份、社保情况、疾病类别和诊断类别等，其拟合后的评价参数见表 8。各变量解释的方差及统计检验结果见表 9。具体参数估计见表 10。

根据表 9 方差分析结果可知，有无社保作为指示变量具有统计学意义，其 p 值为 0.04；各疾病诊断组之间的差异有统计学意义，主要诊断和次要诊断之间的差异有统计学意义。模型的 R^2 为 0.06，变异系数为 201.74。拟合结果中，有无社保的拟合参数的 p 值 0.12 为最大。

表 7　　住院费用测算变量类型及水平统计表

变量类别	变量名称	变量类型	变量水平
保单属性	承保时间	定性	1 个水平： 2021 年 1 月 1 日至 2021 年 12 月 31 日一年期
疾病属性	诊断类型	分类变量	2 个水平： 主要诊断 其他诊断
	疾病类型	分类变量	2 个水平： 阿尔茨海默痴呆 血管性痴呆
客户属性	省	分类变量	31 个水平： 北京 …… 不详
	性别	分类变量	3 个水平： 男 女 不详
	年龄	连续变量	
	有无社保	分类变量	2 个水平： 有 无

表 8 模型参数统计表

	自由度	平方和	均方	F 值	Pr > F
模型	35.00	19940805000000.00	569737299847.00	158.97	<.0001
误差	95374.00	341813030000000.00	3583922545.10		
校正合计	95409.00	361753830000000.00			

表 9 方差分析统计表

	自由度	平方和	均方	F 值	Pr > F
年龄	1	7547715100000.00	7547715100000.00	2105.99	<.0001
省份	29	10235783000000.00	352958049000.00	98.48	<.0001
性别	2	1572112300000.00	786056173687.00	219.33	<.0001
有无社保	1	15053758545.00	15053758545.00	4.20	0.04
疾病类别	1	40545454637.00	40545454637.00	11.31	0.00
诊断类别	1	529595368589.00	529595368589.00	147.77	<.0001

表 10 广义线性模型的参数估计值

参数	系数	标准误差	t 值	Pr > |t|
截距	-36699.47	2968.02	-12.36	<.0001
年龄	771.00	18.70	41.22	<.0001
男性	12702.53	2414.35	5.26	<.0001
女性	4546.36	2420.84	1.88	0.06
性别不详	—	—	—	—
社保:无	-753.09	487.24	-1.55	0.12
社保:有	—	—	—	—
阿尔茨海默痴呆	1615.36	411.27	3.93	<.0001
血管性痴呆	—	—	—	—
主要诊断	-5929.82	487.81	-12.16	<.0001
其他诊断	—	—	—	—

在 GLM 的模型拟合中,省份的参数估计值对模型的结果影响较大。参数估计值绝对值较大的省份包括吉林、安徽、辽宁、陕西等,这些省份的参数估计的 p 值均是显著的。此外,从表中还可以看到,男性和女性的住院费用高于性别不明的分组。年龄作为连续变量进入模型,其参数估计值为 771.00,$p < 0.0001$,标准误为 18.70。说明年龄与住院费用正相关,即随着年龄的增长,住院费用有增高的趋势。

(二)机器学习模型结果展示

随机森林、LASSO 回归和 LightGBM 三种机器学习模型拟合结果见表 11。其中,LightGBM 的均方根误差最小为 28252.47,且 R^2 最大,为 0.27。LASSO 回归的均方根误差最大为 32053.12,且 R^2 最小,为 0.14。

随机森林还可以给出各个因子对拟合结果影响的大小排序,见图 2。从图中可以看到:对住院费用影响最大的是年龄,远高于其他因子;对费用拟合结果影响前三位的因素分别是年龄、是否有医保和性别。此外,随机森林模型还能在调整不同疾病分组的并发症因素后,再次对不同并发症对费用的影响大小进行排序,结果见图 3。从图中可以看到:

考虑并发症因素后，对住院费用影响最大的仍是年龄，但影响第二大的因素变成了并发症，是否有医保的影响排在第三位。最后调整并发症因素后，随机森林模型还给出了不同并发症对费用的影响程度。

表 11　　机器学习模型拟合结果

	RMSE	R^2	MAE
随机森林	28273.61	0.25	17901.20
LASSO 回归	32053.12	0.14	18034.15
LightGBM	28252.47	0.27	17803.41

（三）模型对比分析

模型对比分析基于平均绝对误差和均方根误差，其中平均绝对误差 MAE（Mean Absolute Error，MAE）是绝对误差的平均值，其实是更一般形式的误差平均值，其表达式为：

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (4)$$

而均方根误差 RMSE（Root Mean Squared Error，RMSE），有时称为 RMSD，它可以测量误差的平均大小，定义为预测值和实际观测之间平方差异平均值的平方根，其表达式为：

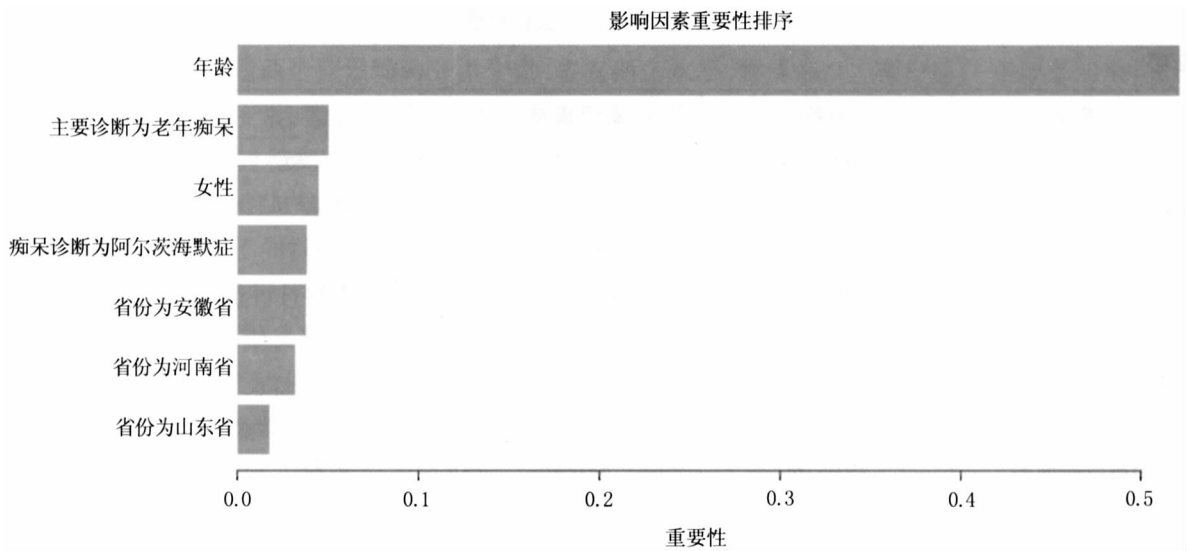

图 2　随机森林影响因素大小排序（调整并发症前）

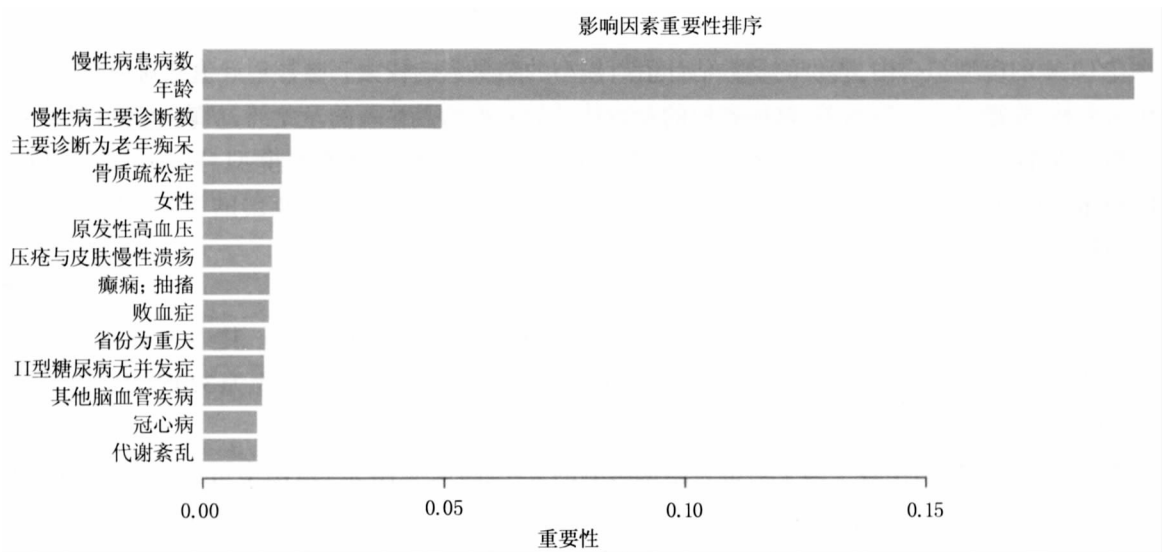

图 3　随机森林影响因素大小排序（调整并发症后）

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_1 - y_t)^2}{T}} \quad (5)$$

GLM 传统模型方法和机器学习模型的参数对比见表 12。其中,LightGBM 模型的 R^2 最大,且均方根误差较小,由此可以认为 LightGBM 模型的拟合效果最好;但随机森林模型可以调整并发症种类,并给出各因素影响程度的相对大小,其 R^2 与 LightGBM 接近,为 0.25,均远高于 GLM 模型的 R^2。

表 12　　模型拟合效果对比

	RMSE	R^2	MAE
随机森林	28273.61	0.25	17901.20
LASSO 回归	32053.12	0.14	18034.15
LightGBM	28252.47	0.27	17803.41
GLM	29674.91	0.06	19865.87

综合以上结果可以看出,在老年痴呆症患者的未来住院费用预测方面,机器学习效果优于传统保费损失厘定使用的 GLM 模型。选择合适的机器学习模型拟合能够获得令人满意的效果。具体来说,在用于损失厘定时机器学习相对于传统的 GLM 模型具有一定优势,通常对数据的分布要求更低,使用更方便,适用范围更广,效率更高;随机森林模型适用于维度不太高,但对准确性有要求的数据,但是容易在数据集相对小或者是低维数据集的时候产生过拟合,计算速度比单个的决策树慢,在推断超出范围的独立变量或非独立变量时效果不佳;而 LASSO 回归在解决数据集中各变量存在共线性问题的情况时效果较好。从本次研究的结果来看,数据集各变量之间的共线性问题影响不大。上述结果还显示,LightGBM 在传统机器学习 GBDT 的基础上,大幅度提高了计算效率,在保证效率的同时提高了拟合效果的精准度。

(四)机器学习模型费用预测

最终的预测结果见表 13,该结果是利用拟合效果较好的 LightGBM 机器学习模型得出的。该结果与既往研究报道的老年患者住院费用的变化趋势结构是相一致的(严敬琴,2019;黄茂娟等,2017;郑金坡等,2017)。

(五)纯保费测算

机器学习模型给出的住院费用测算和估计可以在保险业务中的多个场景应用,例如保费测算、客户服务和风险控制等。这里以纯保费测算为例进行说明。纯保费测算考虑的定价因素是年龄组(60~85 岁,每 5 岁一组)、免赔额(2000 元、5000 元两种情况)、以及个人自付比例(自付 20%、自付 10%、无自付三种情况),其中个人自付比例反映了医保负担的情况。表 14 描述了根据华北地区某省数据所测算的痴呆症带病投保一年期百万医疗险的纯保费情况。结果显示,当免赔额为 2000 时,纯保费较高,但当免赔额提升到 5000 元时,纯保费大幅度下降,具有较好的市场潜力。

表 13　　机器学习 LightGBM 模型住院费用预测

年龄组	总费用	药品费用	占比	耗材费用	占比	医保内费用	占比
全体患者	29352.34	16331.23	56%	5467.86	19%	24509.20	83.50%
50~54 岁	15263.21	8398.51	55%	3411.94	22%	11024.62	72.23%
55~59 岁	16942.17	9236.84	55%	3471.65	20%	12059.44	71.18%
60~64 岁	19314.07	10529.99	55%	3705.83	19%	13929.31	72.12%
65~69 岁	22018.04	11854.59	54%	4101.60	19%	16264.73	73.87%
70~74 岁	25100.57	13514.23	54%	4582.31	18%	19272.22	76.78%
75~79 岁	28614.65	15920.79	56%	5117.22	18%	22782.98	79.62%
80 岁及以上	32287.57	19678.44	61%	5533.47	17%	28225.79	87.42%

表 14　　痴呆症带病投保一年期百万医疗险纯保费情况——以华北地区某省为例

年龄	免赔额	自付比例	保费	年龄	免赔额	自付比例	保费
60～64	2000	20%	1973.14	75～79	2000	20%	1541.48
	2000	10%	2219.78		2000	10%	1734.16
	2000	0%	2466.42		2000	0%	1926.85
	5000	20%	774.48		5000	20%	643.70
	5000	10%	871.28		5000	10%	724.16
	5000	0%	968.09		5000	0%	804.62
65～69	2000	20%	1702.65	80～85	2000	20%	1319.85
	2000	10%	1915.48		2000	10%	1484.84
	2000	0%	2128.31		2000	0%	1649.82
	5000	20%	664.13		5000	20%	516.82
	5000	10%	747.14		5000	10%	581.42
	5000	0%	830.16		5000	0%	646.02
70～74	2000	20%	1639.78				
	2000	10%	1844.75				
	2000	0%	2049.73				
	5000	20%	715.80				
	5000	10%	805.28				
	5000	0%	894.75				

五、结语

本研究充分利用样本量较大，覆盖年龄范围广，研究对象在时间和空间上分布均匀的数据，并通过与传统的 GLM 模型对比，证实了机器学习方法能够较为准确地为带病体住院费用进行预测，为带病投保的保费测算提供基础，为百万医疗险带病投保产品的设计创新提供依据。研究结果可以帮助保险公司扩大投保人群，创新保险产品，为已经患病的人群提供经济保障，从而切实解决老年人的医疗需求，助力健康中国，做到应保尽保，减轻疾病医疗费用负担。

参考文献：

[1]安磊，张洁，齐霞，等. 基于随机森林输变电线路工程造价估算研究[J]. 控制工程. 2016, 23(11): 1841－1844.

[2]陈熙，张晓博. 基于 LightGBM 的住房租金预测分析[J]. 产业与科技论坛. 2020, 19(6): 103－105.

[3]董斌. 百万医疗保险产品设计研究[D]. 深圳大学, 2018.

[4]冯菁楠，王胜锋，詹思延. 基于医疗保险数据的数据库准确性验证方法学进展[J]. 中华流行病学杂志. 2019, (10): 1324－1328.

[5]韩耀风，覃文峰，陈炜，等. adaptive LASSO logistic 回归模型应用于老年人养老意愿影响因素研究的探讨[J]. 中国卫生统计. 2017, 34(1): 18－22.

[6]韩颖，田立启，张云，等. 老年痴呆患者直接经济负担影响因素分析[J]. 中国公共卫生管理. 2016, 32(3): 321－323.

[7]韩颖. 老年痴呆住院患者疾病经济负担及影响因素研究[D]. 青岛大学, 2017.

[8]郝君. 基于随机森林的非寿险准备金索赔次数预测研究[D]. 天津财经大学, 2018.

[9]黄茂娟，潘敏，吴颖敏，等. 四川省老年慢性病患者住院费用构成及影响因素分析[J]. 卫生软科学. 2017, 31(1): 15－19.

[10]贾延延,冯键. 机器学习算法保险场景应用[J]. 合作经济与科技. 2020,(9):132 - 133.

[11]贾友波,宋宪锟. 医院病案疾病诊断 ICD 编码准确性影响因素分析与处理对策研究[J]. 当代医学. 2020,26(6):117 - 120.

[12]康昊昱. 上海市商业长期护理保险发展模式研究[D]. 复旦大学,2010.

[13]雷婷. 苏州市接受机构护理的老年期痴呆的疾病经济负担及影响因素研究[D]. 苏州大学,2011.

[14]李红梅,吴喜之,王涛. 基于纵向数据与多重共线性数据的神经网络与传统方法比较[J]. 统计与决策. 2020,(9):22 - 25.

[15]李阳,陈晓泓,王一梅,等. 基于 LASSO 变量选择联合贝叶斯网络构建恶性肿瘤相关急性肾损伤(AKI)风险预测模型的研究[J]. 复旦学报(医学版). 2020:1 - 10.

[16]梁志胜. 医疗保险对老年医疗服务和健康影响研究[D]. 广西医科大学,2017.

[17]林鹏程,唐辉. 一种改进 Deep Forest 算法在保险购买预测场景中的应用研究[J]. 现代信息科技. 2019,3(22):116 - 122.

[18]吕国友,何月枝,郑美莲,等. ICD - 10 编码检索系统应用分析[J]. 世界最新医学信息文摘. 2019,19(45):34 - 35.

[19]马晓君,沙靖岚,牛雪琪. 基于 LightGBM 算法的 P2P 项目信用评级模型的设计及应用[J]. 数量经济技术经济研究. 2018,35(5):144 - 160.

[20]孟生旺,黄一凡. 驾驶行为保险的风险预测模型研究[J]. 保险研究. 2018,(8):21 - 34.

[21]孟生旺,李天博,高光远. 基于机器学习算法的车险索赔概率与累积赔款预测[J]. 保险研究. 2017,(10):42 - 53.

[22]孟生旺. 神经网络模型与车险索赔频率预测[J]. 统计研究. 2012,29(3):22 - 26.

[23]牛犇. 年龄相关性疾病的 ICD - 10 编码分析[J]. 中国病案. 2017,18(12):41 - 43.

[24]潘兴. 我国商业健康保险风险管理研究[D]. 对外经济贸易大学,2014.

[25]舒晓燕. 护理干预提高老年痴呆症患者的就医依从性[J]. 中外妇儿健康. 2011,19(8):380.

[26]孙东雅,张铭哲. 百万医疗险发展与监管[J]. 中国金融. 2019,(22):68 - 69.

[27]王硕. 我国商业健康险市场发展研究[D]. 湖南大学,2018.

[28]王文文. 基于支持向量机的住院费用预测模型研究[D]. 宁夏医科大学,2016.

[29]王莹,刘克新,林海丽. 老年痴呆疾病编码的回顾性分析[J]. 中国病案. 2006,(7):29 - 30.

[30]王悦,王延博,王辛格. 基于 LightGBM 的乳腺癌预测模型[J]. 智慧健康. 2019,5(29):39 - 41.

[31]吴绍武,续育茹. 基于 LightGBM 的血压检测方法研究[J]. 生物医学工程研究. 2019,38(3):312 - 315.

[32]夏涛,徐辉煌,郑建立. 基于机器学习的冠心病住院费用预测研究[J]. 智能计算机与应用. 2019,9(5):35 - 39.

[33]徐贝尔. 广州市基本医疗保险公平性研究[D]. 华南理工大学,2016.

[34]许译芝. 基于广义线性模型的车联网保险费率厘定研究[D]. 山东大学,2019.

[35]严敬琴. 深圳市某精神专科医院 2013 - 2017 年住院费用构成及影响因素分析[D]. 安徽医科大学,2019.

[36]曾宇哲,吴媛博,郑宏远,等. 基于机器学习的车险索赔频率预测[J]. 统计与信息论坛. 2019,34(5):69 - 78.

[37]张碧怡,肖宇谷,曾宇哲. 车险定价中风险因子重要性测度的比较研究——基于集成学习方法和广义线性回归模型[J]. 保险研究. 2019(10):73 - 83.

[38]张亦鼎. 基于车联网数据的车辆驾驶画像分析和风险研究[D]. 上海交通大学,2019.

[39]郑金坡,马利,徐雅,等. 新疆兵团某三甲医院职工医保老年人住院情况及费用[J]. 中国老年学杂志. 2017,37(18):4646 - 4648.

[40]中央政府门户网站. 我国“重大疾病保险疾病定义使用规范”正式启用[Z]. 2007:2020.

[41]Huang J, Tsai Y, Wu P, et al. Predictive Modeling of Blood Pressure During Hemodialysis: A Comparison of Linear Model, Random Forest, Support Vector Regression, XGBoost, LASSO Regression and Ensemble Method[J]. Elsevier B. V. 2020.

[42]Jia J, Wei C, Chen S, et al. The Cost of Alzheimer's Disease in China and Re - Estimation of Costs Worldwide[J]. Alzheimer's & Dementia. 2018, 14(4):483 - 491.

[43]Kelley A S, Mcgarry K, Gorges R, et al. The Burden of Health Care Costs for Patients with Dementia in the Last

5 Years of Life[J]. Annals of Internal Medicine. 2015, 163(10):729.

[44] Moon S, Seo H, Lee D Y, et al. Associations among Health Insurance Type, Cardiovascular Risk Factors, and the Risk of Dementia: A Prospective Cohort Study in Korea [J]. International Journal of Environmental Research and Public Health. 2019, 16(14):2616.

[45] Nelder J A, Wedderburn R W M. Generalized Linear Models[J]. Journal of the Royal Statistical Society. 1972, 135(3):370 - 384.

[46] World Health Organization. Dementia[Z]. 2019: 2020 https://www.who.int/news - room/fact - sheets/detail/dementia.

[47] Pyenson B, Sawhney T G, Steffens C, et al. The Real - World Medicare Costs of Alzheimer Disease: Considerations for Policy and Care[J]. J Manag Care Spec Phann. 2019, 25(7):800 - 809.

[48] Tihshirani R. Regression Shrinkage and Selection via the Lasso: A Retrospective[J]. Journal of the Royal Statistical Society. Series B(Statistical Methodology). 2011, 73(3):273 - 282.

[49] Wang S, Cheung D S K, Leung A Y M. Overview of Dementia Care under the Three - Tier Long - Term Care System of China[J]. Public Health Nursing. 2018.

[50] Xu J, Wang J, Wimo A, et al. The Economic Burden of Dementia in China, 1990 - 2030: Implications for Health Policy[J]. Bulletin of the World Health Organization. 2017, 95(1):18 - 26.

[51] Zhu Y, Liu H, Lu X, et al. Prevalence of Dementia in the People's Republic of China from 1985 to 2015: A Systematic Review and Meta - Regression Analysis[J]. BMC Public Health. 2019, 19(1).

Projection and Comparison of the Hospitalization Expenses of Medical Insurance Covering Alzheimer's Disease: Based on the Machine Learning Model

Zhao Yingxu Bao Zhuqing Gao Shan Zhou Liang
Liu Yisheng Chen Hao Zhang Ning

Abstract: Alzheimer's disease seriously affects the quality of life of the elderly, and its impacts continue to deepen as the proportion of the aged keeps rising and the life expectancy increases in China. The development of commercial health insurance for Alzheimer's disease becomes more and more important. Based on detailed historical data, the paper used the generalized linear model to estimate pricing factors to predict the hospitalization costs of Alzheimer's patients. Meanwhile, it also applied the machine learning methods like random forest, LASSO regression and Light GBM to estimate and predict hospitalization costs of Alzheimer's patients. This research is of great significance for insurance companies to develop diversified products, enrich the insurance market, and help the elderly. The horizontal comparison of the four methods show that machine learning methods have greater advantages in cost estimation, and also have greater potential in the number of influencing factors, precise pricing and model adaptation. The final calculation results and the process method combined with the hospitalization rate data can be used to specifically calculate the medical insurance cost of Alzheimer's disease.

Key words: Alzheimer's disease; commercial medical insurance; applying for insurance with existing disease; machine learning; premium calculation