

基于双层预测模型的新能源车险定价研究

张毓科

(湖南大学金融与统计学院, 湖南省长沙市, 410006)

摘要: 大力发展新能源汽车是我国推动绿色发展, 建设美丽中国的重要举措。近年来新能源汽车的保有量快速增加, 但在新能源车险定价中存在定价因子多, 定价模型需要改进的问题。本文创新地提出了一个双层模型对新能源车险进行定价, 该双层模型基于广义线性模型和 XGBoost 算法。使用广义线性模型作为底层输出可以保证模型的可解释性, 加入 XGBoost 算法之后又可以增加模型的预测精度, 因此双层预测模型具有良好的性能。基于一组实际的新能源汽车理赔数据进行了实证分析, 实证结果表明双层预测模型的预测效果更好, 贝叶斯信息准则和模型偏差度均小于单一的广义线性模型, 并且双层模型能更好地预测极端损失。

关键词: 新能源车险; 广义线性模型; XGBoost 算法; 双层预测模型

中图分类号: F8

文献标识码: A

1. 引言

党的二十大报告中指出: 我们要加快绿色转型发展, 倡导绿色消费, 推动形成绿色低碳的生产生活方式。新能源汽车的普及是我国实现高质量发展、建设美丽中国的必由之路, 对此国家陆续出台了一系列支持新能源汽车发展的政策, 如购买新能源汽车可以免除购置税、充电设施运营单位可以申请补贴等等。新能源汽车在我国发展迅速, 据公安部的统计数据显示, 2022 年我国新能源汽车的保有量达到了 1310 万辆, 同比增长率高达 67%。此外, 消费者对于新能源车的接受程度也越来越高, 2022 年新能源汽车的销量达到了 688.7 万辆, 同比增长 93.4%。在政策利好和市场利好的情况下, 新能源汽车持续快速发展, 具有良好的发展空间。汽车保险是汽车的产后市场, 2021 年 12 月, 中国保险行业协会发布了《中国保险行业协会新能源汽车商业保险专属条款(试行)》, 该条款明确了新能源车险的保障范围, 对电池及储能系统、电机及驱动系统和其他控制系统提供了保障, 并增加了充电桩等附加险。2022 年新能源车险保费规模突破 700 亿元, 增长率超过 100%, 随着新能源汽车的快速增长, 新能源车险将给车险市场带来持久的增长动力。

费率厘定因子多, 损失预测模型需要改进是当前新能源车险定价中存在的主要问题。新能源汽车在结构组成上与燃油车有很大差异, 因此在车险定价中需要选取的定价因子存在不同, 电池类型、驱动方式等核心变量应当纳入到新能源车险的定价中去。此外, 随着定价因子的增多, 相应的定价模型也要进行改进。为了更好地拟合新能源车险的风险损失, 本文在定价因子和定价模型两方面都进行了改进。在定价因子方面, 本文加入了电池材料、驱动方式等新能源汽车特有的变量和行驶里程、行驶时长等驾驶行为动态变量。在定价模型方面, 本文创新性地提出了一个基于广义线性模型和 XGBoost 算法相结合的双层预测模型, 该模型可以在兼顾可解释性的基础上提高模型的预测精度。

2. 文献综述

2.1 车险定价因子

车险定价是精算领域研究的热点问题, 传统车险定价所采取的定价因子为年龄、性别等驾驶人特征和车辆品牌、车龄等车辆属性, 新能源车险与传统机动车的核心区别在于驱动方式不同, 因此在定价中要考虑新能源汽车所特有的变量。许闲(2019)指出电池是新能源汽车的关键部件, 但由于电池成本高、技术复杂等原因, 精细化的产品定价较为困难, 建议应当完善电池风险评估体系, 建立合理的新能源汽车费率厘定系统^[1]。邵鲁扉(2019)对新能

源汽车的理赔数据进行单因素分析并建立广义线性模型,结果都表明电池的生产厂商和型号等信息对于风险的区分具有重要的作用,并提议要收集电池使用频率,充电状况等动态信息[2]。

随着大数据和车联网的快速发展,可用于车险定价的因子大量增加。通过车载设备可以收集驾驶人的驾驶行为信息,将其加入到定价模型中可以更加充分地对投保人的风险进行评估,实现精细化的定价。韩成卉,赵绰翔,郑苏晋(2015)将人的行为和心理等因素放进了车险定价模型中,提出了“行为保险定价”,为车险费率厘定因子的选择和无赔款优待系统的调整提供了理论基础[3]。Baecke P, Bocca L (2017) 研究表明通过物联网大数据能够帮助保险公司更好的进行风险评估,根据驾驶情况付费的相关变量是最有价值的风险因子,且这种基于远程信息的保险产品可以快速实施。此外,通过车辆测试实验可以获得更详细的汽车参数,加入定价模型中可以提高模型的预测精度[4]。张琳,黎星言(2022)将中国保险汽车安全指数加入到费率厘定模型中,结果表明在车损险的定价中,加入维修经济性得分和碰撞性得分可以提高模型的拟合效果[5]。

2.2 车险定价模型

关于车险定价模型的研究可以分为两个方向,一是使用经典的广义线性模型,或者在广义线性模型的基础上,通过改变统计假设以构造出衍生模型;二是使用决策树、神经网络等机器学习模型。

广义线性模型(Generalized Linear Model, GLM)是当前车险定价研究中所使用的主流模型,孟生旺(2007)系统阐述了广义线性模型的结构和优势,并对一组第三者责任险的理赔数据进行了预测[6]。由于被保人的熟练程度、车辆性能等存在区别,实务中的损失数据呈现出零膨胀、异质性、长尾性等特征,单一分布的广义线性模型不能很好地拟合,之后大量学者在传统 GLM 的基础上进行改进,对因变量的分布特征进行扩展,构建出多种 GLM 的衍生模型,以更加符合现实世界的特征。孙维伟,张连增,胡祥(2017)提出的分层广义线性模型可以处理纵向数据与分层结构数据,拟合的效果更好[7]。此外,在模型设定中纯风险损失可由索赔频率与索赔强度的乘积得出,也可直接对纯风险损失进行预测,张连增,谢厚谊(2017)对比了直接对纯保费建模与分别对索赔频率和索赔强度建模两种方法,结果表明从待估参数的角度看,直接使用 Tweedie-GLM 模型对纯保费进行预测效果更好[8]。

近年来,随着机器学习和人工智能的飞速发展,利用机器学习模型进行保险损失的预测成为研究的热点。使用机器学习模型可同时加入大量的费率因子,增强预测结果的准确性。Liu, Wang, Lv (2014) 采用 Adaboost 算法对索赔强度进行预测,并将结果与 GLM、神经网络进行对比,结果表明 Adaboost 算法表现良好,具有最高的预测精度和相对较小的方差[9]。张连增,谢厚谊(2018)基于泊松回归树对索赔频率进行建模,并与广义线性模型的结果进行比较,结果表示 Bagging 法使得回归树的预测精度和 GLM 相当,但在解释性上不如广义线性模型[10]。Huang, Meng (2019) 将逻辑回归、支持向量机、随机森林、XGBoost 和人工神经网络作为风险概率模型,构建了具有显著预测效果的分层费率厘定模型。此外,有文献将广义线性模型与机器学习进行结合,以达到更好的预测效果[11]; Maillart (2021) 基于树集合的方法构建了车险的索赔频率预测模型,并把这种方法结合到广义线性模型中[12]。张连增,江璐嘉(2022)提出了一种基于进化树的数据分箱方法,可以将连续型变量划分为离散型,将划分之后的数据与广义线性模型相结合,可以达到较好的预测效果,且保证了模型的简便性和可解释性[13]。

通过对相关文献的梳理可以发现,目前使用机器学习算法进行精算预测大多都集中在单个模型,侧重于单个算法之间的比较,以寻找最优的预测模型。将机器学习算法与传统的精算模型相结合,进行复合模型的研究较少。

3. 模型建立

为综合广义线性模型的可解释性和机器学习算法的预测准确性,本文构建了双层费率厘

定模型。首先使用广义线性模型对纯风险损失进行预测，使用重要的从人因子和从车因子变量，作为第一层模型；然后使用机器学习算法对第一层模型的残差进行预测，可使用动态驾驶行为因子等多个变量，生成残差的预测值；最后将预测的残差因子作为一个新的变量，与第一层模型中使用的变量结合，再次对反应变量进行预测，得到最终的预测结果。模型结构如图 1 所示。

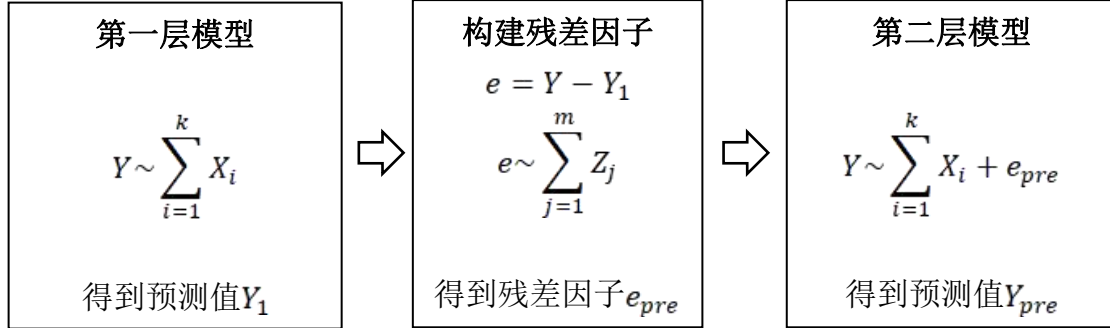


图 1 双层预测模型结构图

3.1 Tweedie-GLM

在双层预测模型中的第一层和第二层均使用了 Tweedie-GLM 模型，直接对纯风险损失进行建模。模型结构如式(1)所示。

$$\begin{cases} Y \sim \text{Tweedie}_p(\theta, \varphi) \\ g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \end{cases} \quad (1)$$

在本文构建的广义线性模型中，实际损失值 Y 为随机部分，由于很多保单并没有出险，索赔额在 0 处有大量堆积，因此假设其服从 Tweedie 分布。式(1)中 μ 表示风险损失的均值， $g(\mu)$ 为连接函数，本文采取了对数连接函数； η 为系统部分，在第一层模型中只使用了从人和从车的定价因子，在第二层输出模型中加入了构建的残差因子。

此外，在进行分类费率厘定时，要对变量进行分组处理(孟生旺(2007))。将连续型变量进行分组的意义在于可以保证每个区间数据量的充足性。比如对于车龄变量，如果每个车龄段单独作为一组，则每个组别之间的样本数量可能不足，相近车龄段所对应的纯风险损失比较接近，将其合并为一组可以保证该区间具有统计学意义，满足保险费率厘定中大数法则的要求。关于对连续变量进行分组的方式，本文采取了张连增等(2022)提出的基于数据驱动的分箱方式。具体的分组流程为：首先确定需要划分的区间数，其次将真实的损失额作为因变量，需要划分区间的连续变量作为自变量，设置划分区间数为最大叶子节点数，最后使用回归决策树进行拟合，可将连续变量划分为分类型变量。

3.2 XGBoost 算法预测残差

第一层构建的广义线性模型生成预测值后，中间层模型采用机器学习算法对残差进行预测。采用机器学习算法进行拟合的优点有两个，一是可以挖掘费率厘定因子对于反应变量的非线性关系；二是利用机器学习算法可以不考虑自变量的数量，进行高维度数据的拟合，也不需要考虑因变量的统计学假设分布。

对残差进行预测的核心思想在于对误差进行预测拟合，再返补到上一次的预测中，将多次预测拟合的结果综合起来，以达到较好的拟合效果。Jerome(2001)基于回归树进行了特殊的加强，并指出梯度增强的回归树是准确性高且稳健的^[14]，因此在对残差进行预测时选择了梯度下降树模型，具体而言，本文选择了 XGBoost 算法进行了残差的预测，一是由于构建双层模型的核心思想就是对残差进行拟合并加入到原有的费率厘定因子中，这与 XGBoost 逐步对树的误差进行拟合的思想相符合；二是由于 XGBoost 算法对目标损失进行了二阶泰勒展开，能够达到较高的预测效果，并且加入了系数的惩罚项，可以得出较为稳定的结果，XGBoost 算法的目标损失函数如式(2)所示。

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

3.3 模型优势分析

本文将广义线性模型与 XGBoost 算法相结合，使用双层模型进行新能源车险的定价研究，模型具有以下三点优势。

一是有效提高模型的预测准确度。由于对残差进行了拟合，在第二层的模型中加入了残差因子，可以有效地弥补预测的准确性。使用 XGBoost 模型对残差进行预测可以不用考虑自变量的数量，并且可以使用驾驶人行为因子等多维度数据。

二是保证模型具有较好的可解释性。由于第二层模型依然为广义线性模型，最终预测值的输出是以广义线性模型的形式，因此模型具有良好的解释性，一个合理的模型需要有较强的解释能力，模型结果要明确知道各个风险因子和赔付之间的关系是什么，从而能够帮助到一线承保的业务人员，透明的模型能给业务带来更多的“安全感”和“信任感”。

三是加入残差因子后不会影响重要变量的输出解释能力。第一层使用广义线性模型拟合后，因变量对自变量的线性关系基本都已经体现在系数中，因此残差中包含着未纳入模型的遗漏变量的影响和模型中变量的非线性影响。在对残差进行拟合时，也可以加入模型中原有的变量，这样可以进一步挖掘原有自变量对纯风险损失的非线性影响。将拟合的残差因子加入到原有费率厘定模型进行第二层模型时，由于残差因子所包含的信息基本为非线性影响，因此原始的费率厘定因子系数并不会有很大的变化。在实务中可以把重要的费率厘定因子单独放入广义线性模型中，进行分组并求得系数，对于重要性不强且数量很多的变量则可以放到残差因子中，提高模型整体的预测能力。

4. 实证分析

4.1 数据说明

本文所使用的数据为某大型财险公司的新能源车理赔数据，经过数据的清洗和变量的初步筛选后，该数据包含了 96207 条保单，每条保单记录有 21 个变量，包括损失和费率厘定因子，定价因子分为传统从人从车的静态定价因子和驾驶行为因子，具体的数据格式如表 1 所示。

表 1 实证分析数据格式

索赔额	动力类型	车龄	组合投保方式	车辆品牌	电池材料	电池包总容量	日均行驶里程	...
0	插电式混合动力	0	主全	比亚迪牌	三元材料	37	18.44	
2156.11	电动	1.94	交三	江淮牌	磷酸铁锂	85	46.38	...
0	电动	0	单交	宝骏牌	磷酸铁锂	131	20.28	
...								

索赔额为本次建模的反应变量，实际损失额为 0 的值有 78786 个，有超过八成的保单并没有发生索赔，数据的索赔率为 18.11%。数据中最大的索赔额为 253666，所有保单平均索赔额为 2364。费率厘定因子的说明见表 2 所示。

表 2 静态定价因子

变量名称	变量类型	变量说明
组合投保方式	分类型	{主全、交三、单交}
车辆品牌	分类型	东风牌等，共 30 个品牌
车龄	数值型	取值为 [0—2.73]（单位：年）
动力类型	分类型	{电动，插电式混合动力}
电池材料	分类型	{三元材料、磷酸铁锂}
电池包总容量	数值型	取值为[24.7—371.2]（单位：ah）

组合投保方式为投保人在购买车险时的组合方式。新能源车险包括法律强制购买的交强险和自愿购买的商业车险,商业车险包括车损、三责和车上人员险等。主全表示投保人购买了所有的车险主险,交三表示购买了交强险和第三者责任险,单交表示只购买了交强险。产生不同组合投保方式的原因可从保险公司和投保人两个角度来看:对于保险公司来讲,将不同险种组合出售是提高保费规模的重要手段,且可以控制成本,一般来讲车损险损失率较大,因此“单保车损险”会被严格把控,会与商业三责险等一起打包出售。对于投保人来讲,险种的组合方式反映了对自身驾驶能力的评价,对于一个驾驶能力自信的司机,可能倾向于只投保交强险,而对于新手司机一般会投保所有主险。因此投保方式可以反映驾驶人的特征,在本文中当作从人的费率厘定因子引入到模型中去。

车辆品牌和车龄为从车方面的定价因子,车龄为0的保单数量为54349,这是由于新能源车在最近几年才逐渐被大众接纳,因此新车占比较高。电池类型、动力类型和电池包总容量是新能源车区别于传统汽车的变量。车龄和电池包总容量为数值型变量,为实行分类费率厘定,需要将该变量进行分箱处理,本文使用回归决策树模型进行了数值驱动型分组方式,将车龄和电池包容量划分为了三组,以索赔额作为反应变量,最大叶子节点数设置为3,车龄的分组为 $([0,0.75], (0.75, 1.9], (1.9, 2.73])$,电池包总容量的分组为 $([24.7, 53.0], (53.0, 114.4], (114.4, 371.2])$ 。

此外,本文还引入了驾驶行为的动态因子,该数据为汽车在投保期间,通过OBD设备等车联网工具记录下的行驶数据,共14个指标,分为了8个维度,均为数值型变量,具体的指标如表3所示。

表3 驾驶行为动态因子

变量维度	变量名称	变量维度	变量名称
日均指标	日均行驶里程	充电情况	日均充电次数
	日均行驶次数		日均充电时长
夜间行驶	夜间行驶时长占比		百公里充电次数
早晚高峰	早晚高峰行驶时长占比		快充次数占比
节假日	节假日行驶时长占比	行驶用电情况	充电起始SOC区间
疲劳驾驶	疲劳驾驶时长占比		舒适温度运行占比

数据说明: SOC 是指电池的荷电状态,英文全称为 State of Charge,指电池中剩余电荷的可用状态,从0-100等分为十个区间(0-10,...,90-100)。

4.2 模型预测

本节构建了两个风险预测模型,分别为基准的广义线性模型和双层预测模型,在双层预测模型中,使用构建的残差因子代替驾驶行为因子,具体结果如表4所示。

表4 模型结果

广义线性模型			双层预测模型	
variable	coefficient	p-values	coefficient	p-values
Intercept	7.4924	0.0000***	5.7760	0.0000***
C(组合投保方式)[T.交三]	-0.8227	0.0000***	-0.8770	0.0000***
C(组合投保方式)[T.单交]	-1.2955	0.0000***	-1.3031	0.0000***
C(车龄区间)[T.B<1.9]	-0.0777	0.1315	-0.0629	0.1411
C(车龄区间)[T.C<2.9]	-0.3581	0.0000***	-0.5061	0.0000***
C(动力类型)[T.电动]	-0.3359	0.2212	-0.3063	0.1805
C(电池类型)[T.磷酸铁锂]	-0.1113	0.0733*	-0.2426	0.0000***
C(电池包间)	-0.0222	0.9336	-0.1618	0.4669
[T.(53.0-114.4)]				
C(电池包区间)	0.3271	0.2309	0.3041	0.1819
[T.(114.4-371.2)]				

残差因子		0.2471	0.0000***
日均行驶里程	0.1555	0.0000***	
日均行程次数	-0.1028	0.0001***	
夜间行驶时长占比	0.0699	0.0034***	
早晚高峰行驶时长占比	-0.0171	0.3797	
节假日行驶时长占比	-0.0092	0.5329	
疲劳行驶时长占比	-0.0515	0.1374	
短行程行驶时长占比	-0.0586	0.0880*	
日均充电次数	0.0876	0.0135**	
日均充电时长	-0.0393	0.1826	
百公里充电次数	-0.0400	0.1066	
快充次数占比	0.0609	0.0043***	
舒适温度运行占比	0.0270	0.3852	
充电起始 soc	-0.0109	0.3778	
行驶起始 soc	-0.0262	0.0049***	
C(车辆品牌)		宝骏牌作为基准	

注：*表示在 10%水平下显著，**表示在 5%水平下显著，***表示在 1%水平下显著

模型结果中的“T”表示对照区间的系数结果，组合投保方式中，单交、交三和主全由于承保的内容增加，因此纯风险损失逐步递增；车龄区间的结果显示车龄越大，纯风险损失越小。主要由两方面因素引起，一是车龄变大之后，旧车相对于新车的维修费用会降低，二是大的车龄伴随着更加熟练的驾驶技术，因此出险的概率也会大大减小。

动力类型变量中插电式混合动力为基准区间，电动区间相对于基准区间的系数为负，说明电动型新能源汽车的纯风险损失小于插电式混合动力，插电式混合动力车的电池相对比较大，能够外部充电，可以用纯电动模式行驶，电池电量耗尽后再以混合动力模式（以内燃机为主）行驶，并适时向电池充电，因此插电式新能源车结构更为复杂，当汽车需要维修的时候，其维修成本要大于纯电动车，其纯风险损失更大；磷酸铁锂电池的系数为负，纯风险损失小于三元材料电池，是由于三元材料电池比磷酸铁锂电池更贵，具有能量密度高、循环性能好和电压平台高等优势，目前在市场上占据主要地位，因此三元材料电池的车辆平均损失更大；不同区间的电池包总容量对应的平均损失没有呈现出单调的变化，先增加后减小。当电池容量过小时，单次续航里程较小，因此充电次数较多，电池发生损坏的可能性增大，因此其纯风险损失较高。当电池容量过大时，发生事故时对应的维修费用较高，索赔强度较大，因此纯风险损失较大。此外，由表 4 可以看出，对于两个不同的风险预测模型，静态定价因子的相对系数基本保持一致，回归结果的正负性没有发生变化，系数值的变化也较小。车辆品牌的模型系数如图 2 所示，基准品牌为宝骏牌。显示不同车辆品牌之间的纯风险损失差异显著，费率厘定时车辆品牌是一个非常重要的厘定因子。

由图 2 可以看出，不同车辆品牌之间系数差异显著，车辆品牌是一个非常重要的从车费率厘定因子，且两个预测模型车辆品牌的系数结果较为相似。由于本文是直接对纯风险损失进行建模，因此造成系数差异的原因有两个，一是由于车辆的指导价格较高，导致出险后维修的成本高，案均赔款大。由于宝骏牌车辆的价格较低，因此基本所有的车辆品牌系数均为正数，回归结果显示宝马牌对应的相对系数最高，是由于其价格高，定位于中高端市场，因此其案均赔款大，纯风险损失较大。二是由于车辆的抗碰撞性、刹车效率等车辆性能导致出险频率不一致，出险次数高的汽车其对应的纯风险损失也会变大，哪吒牌和宝骏牌汽车的平均价格相差不大，但系数出现如此大的差异，是由于数据中哪吒牌的出险频率高。此外，不同系别的车可能具有不同的理念，会导致汽车生产时侧重点不同，进而出险频率也不同。以德系车和日系车为例，德系车更注重技术理念，而日系车更加注重市场理念，德系车的车身较于日本车更硬，因此在发生事故的时候，日系车发生形变的程度更大，会产生更多的理赔，

这是日系丰田牌汽车系数大于德系大众的一部分原因。

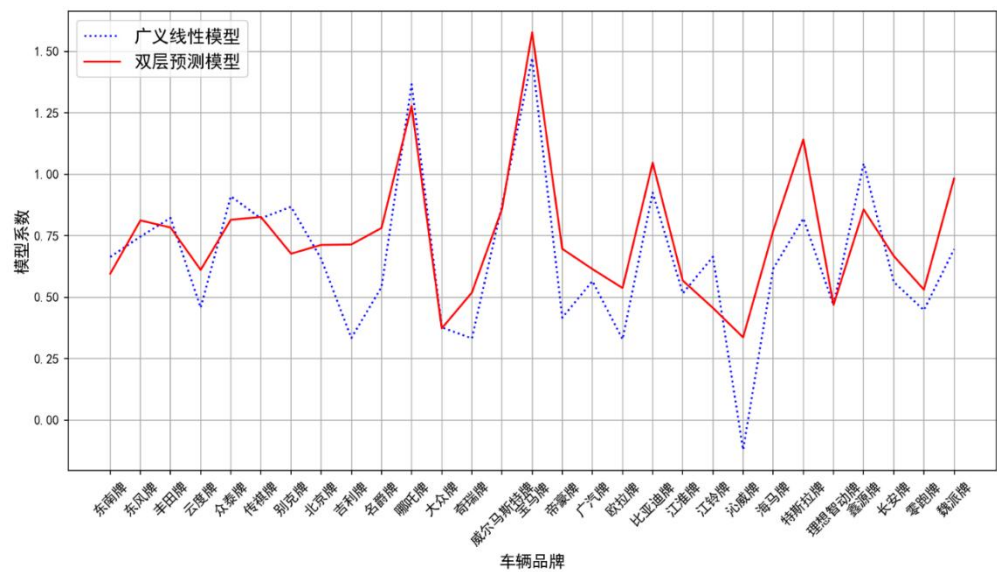


图 2 车辆品牌系数图

由表 4 可知残差因子的系数显著不为 0,残差因子对于纯风险损失的预测具有重要作用,但进行解释时并不能很好的解释残差因子的含义,因为残差因子中包含着驾驶行为因子多个变量,因此在实际应用中加入残差因子的双层模型仅拥有提升模型预测能力的作用,对于需要单独考虑分组的重要解释变量,如车辆品牌,组合投保方式等,则应该单独作为一个变量,不应放入残差的预测中。使用残差因子的好处在于 XGBoost 算法不考虑变量的数量,使用多个变量对残差进行预测可以尽可能利用已有的信息对预测效果进行提升。

在广义线性模型的回归结果中,观测期内行驶里程结果显著,表明行驶里程越大,则对应的纯风险损失越大,基于里程的定价是十分合理的;观测期内日均行程数越大,则对应的纯风险损失越小,在固定的里程下,行驶的次数越多,则单次里程数较小,时间也较短,出险频率下降;夜间行驶占比越大,则对应的纯风险损失越大,其影响机制可能与疲劳驾驶有关,夜间行驶更容易使驾驶人产生疲劳,夜间行驶占比和疲劳驾驶的相关性达到 0.6,两个变量同时进入模型时疲劳驾驶变的不显著,而夜间行驶占比显著为正;短行程行驶时长占比越大,纯风险损失越小,短行程距离短,对应的出险频率较小;日均充电次数越多,快充次数占比越大的保单对应的纯风险损失更大,体现在对电池的损耗上,电池是新能源汽车的核心部件,电池充放电的频率越高,则越容易产生损毁,增加索赔频率;行驶起始的 SOC 区间越大,则纯风险损失越小,行驶起始的 SOC 区间表示行驶起始时的存电量,行驶起始 SOC 区间较小的保单说明在行程开始时电池不是满电状态,在之后的运行过程中可能由于电力不足而产生各种事故,加大的索赔发生的频率,而充电起始 SOC 区间仅表示了充电开始时电池的存量,充电之后不一定马上使用,因此该变量并不显著。

4.3 模型对比

表 5 结果表明,使用 XGBoost 算法对残差进行预测,再使用残差因子构建的双层模型预测效果较好,贝叶斯信息准则和偏差较广义线性模型有显著下降,BIC 值相较于广义线性模型减低约 7%,显著提高了模型的预测能力。

表 5 模型统计指标对比

	贝叶斯信息准则(BIC)	偏差(deviance)
广义线性模型	5358868	6462175
双层预测模型	4965535	6068992

对纯风险损失预测模型的结果进行评价有多种方法,使用统计学的指标去衡量模型的优

劣并不直观，比如模型一的 BIC 比模型二小，说明模型一的预测效果优于模型二，但在实务中的指导意义并不明显，因此本文引入了实务中常用的提升曲线评价指标。该指标并不关注预测出的损失与真实损失之间的绝对差额，而是从风险等级分类的角度对序数效应进行衡量，比如对于两个保单的预测值，保单分组一大于分组二，说明保单分组一的风险等级是高于分组二的，如果模型预测合理，那么保单分组一的真实损失应该大于保单分组二，如果预测出的损失和真实损失的排序情况是一致的，那么说明模型预测的结果是合理的，构建提升曲线的具体步骤如下：

(1) 将保单的风险损失预测值从小到大排序，按照保单数量等分为 n 组，计算每个组内损失预测值的平均值；(2) 计算 n 个组内保单的真实损失的平均值；(3) 以损失预测值的平均值作为横坐标，真实损失额的平均值作为纵坐标，做出提升曲线图。

本文将模型预测出的损失额从小到大排序后，把所有保单等分为 20 组，分别计算各个组内的真实损失额和模型预测出的损失额，构建了两个模型对应的提升曲线。图 3 为广义线性模型的提升曲线，提升曲线整体呈现上升的趋势，表示模型的预测效果整体上是符合实际损失的，但存在几个向下的拐点，说明模型在这几组中出现了风险等级的错配。20 个组是按照预测损失值从小到大的顺序进行排序的，比如在第六组出现了向下的拐点，说明第七组预测的纯风险损失是高于第六组的，但实际的情况却是第七组的真实损失小于第六组，高损失额的保单反而对应的纯保费较低，存在风险等级的错配，预测不准确。

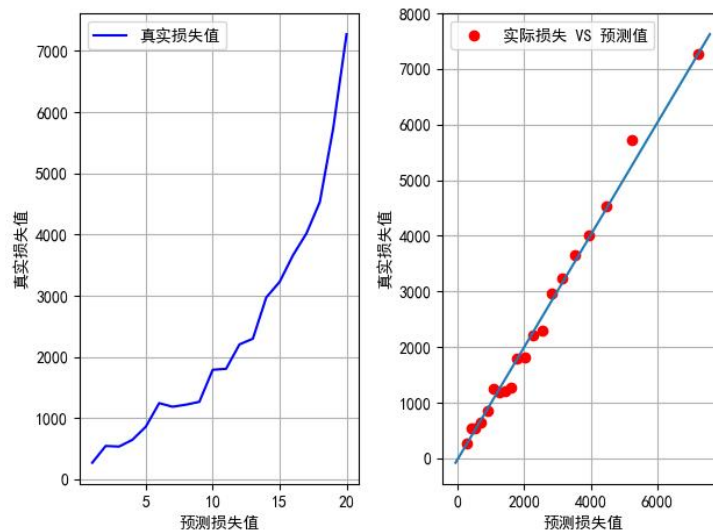


图 3 广义线性模型提升曲线

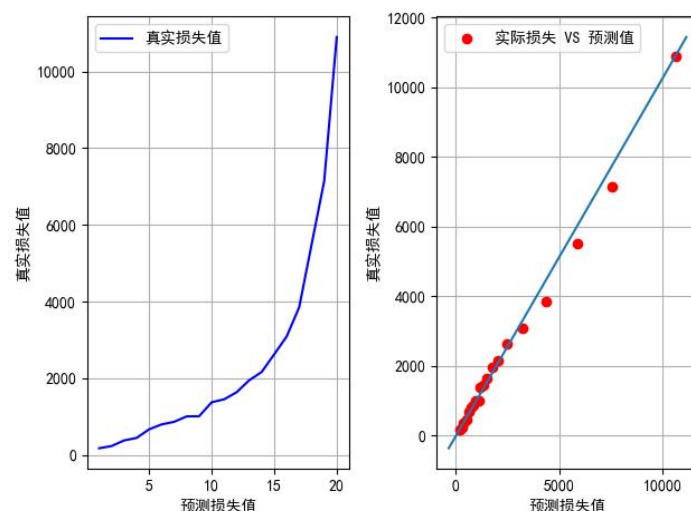


图 4 双层预测模型提升曲线

图4表示双层预测模型的提升曲线更为平缓,基本没有向下的拐点,对于风险等级的预测较为准确。此外,图4的提升曲线表示双层预测模型在极端值的预测值较大,表明双层预测模型可以更好的对极端损失进行拟合,显著优于加入驾驶因子的广义线性模型。对于广义线性模型,第二十组的真实索赔额平均值在7000左右,但对于双层预测模型,第二十组的真实损失值在12000,说明使用双层模型可以增强对极端值的预测能力,使得风险等级更大的保单给予更多的纯保费。

如果提升曲线单调递增的,就说明预测损失大的保单其真实损失也较大,没有风险等级的错配,提升曲线向下的拐点说明预测出的高损失保单其真实损失并不高,存在风险等级的错误预测,模型不合理,因此进一步做出提升度来对提升曲线进行量化。具体构建步骤为:将保单分为20组,则提升曲线共有20个点,计算出两两相邻的点所构成直线的斜率,求出斜率的平均值作为提升度指标。两个模型的提升度如下表6所示,结果表明双层模型的提升度优于广义线性模型,兼顾了模型的预测能力和解释能力,具有很好的预测效果。

表6 模型提升度对比

模型	提升度
广义线性模型	368.6
双层预测模型	564.3

5. 结论

本文利用广义线性模型和XGBoost算法构建了一个双层预测模型,主要结论如下:

第一,电池相关的费率厘定因子是新能源车险定价中重要的变量。实证分析结果表明电动类型的车纯风险损失小于插电式混合动力,磷酸铁锂电池的新能源汽车纯风险损失小于三元材料,电池包容量对于纯风险损失的影响为“倒U型”。第二,本文所构建的残差因子在双层预测模型中作用显著。残差因子的意义在于可以提取费率厘定因子对于损失的非线性影响,且可以同时使用多个变量,因此可以提高模型的预测精度。实证结果表明,残差因子在双层预测模型中系数显著不为零,说明构建的残差因子对于提升模型预测的准确度具有重要作用,且加入残差因子后,原始的从人从车费率厘定因子的系数正负和大小均没有发生大的变化。第三,广义线性模型和XGBoost算法相结合的双层预测模型优于单一的广义线性模型,并且在极端损失的预测上表现较好。首先使用统计指标对构建的两个模型进行对比,双层预测模型的BIC准则和模型偏差均小于广义线性模型。此外,本文还构建了提升曲线对模型进行对比,提升曲线可以衡量模型是否对风险等级进行了错配,结果显示双层预测模型所对应的提升曲线更为平缓,预测效果最好,进而通过提升曲线构建了提升度这一量化指标,结果也表明双层预测模型优于单一的广义线性模型。

参考文献

- [1] 许闲.新能源汽车的风险与保险分析(下)[J].中国保险, 2019(09):21-24.
- [2] 邵鲁康.新能源汽车的保险定价研究[J].中国商论, 2019(12):43-45.
- [3] 韩成卉,赵绰翔,郑苏晋.“行为定价保险”:车险定价与驾驶行为理论[J].金融发展研究, 2015(08):68-74.
- [4] Baecke P, Bocca L. The Value of Vehicle Telematics Data in Insurance Risk Selection Processes[J]. Decision Support Systems, 2017(98):69-79.
- [5] 张琳,黎星言.加入中国保险汽车安全指数指标的车险索赔次数研究[J].汽车工程学报, 2022, 12(03):301-306.
- [6] 孟生旺.广义线性模型在汽车保险定价的应用[J].数理统计与管理, 2007(01):24-29.
- [7] 孙维伟,张连增,胡祥.基于分层广义线性模型的非寿险费率厘定精算模型研究[J].统计与信息论坛, 2017, 32(06):48-54.
- [8] 张连增,谢厚淳.Tweedie分布在车险费率厘定中的应用[J].保险研究, 2017(01):80-90.

- [9] Yue Liu, Bing Jie Wang, Shao Gao Lv. Using Multi-class AdaBoost Tree for Prediction Frequency of Auto Insurance[J]. Journal of Applied Finance & Banking, 2014, Vol.4(5): 45-53.
- [10] 张连增, 谢厚谊. 回归树方法在车险索赔频率预测建模中的应用[J]. 保险研究, 2018(01): 101-111.
- [11] Yifan Huang, Shengwang Meng. Automobile insurance classification ratemaking based on telematics driving data.[J]. Decision Support Systems, 2019, Vol.127: 113156.
- [12] Arthur Maillart. Toward an explainable machine learning model for claim frequency: a use case in car insurance pricing with telematics data[J]. European Actuarial Journal, 2021, 11(2): 579-617.
- [13] 张连增, 江璐嘉. 车险定价中风险保费类别的构造——基于广义线性模型与数据驱动的分箱方法[J]. 中央财经大学学报, 2022(09): 25-38.
- [14] Jerome H Friedman. Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, Vol.29(5): 1189-1232.

Research on new energy vehicle insurance pricing based on two-layer prediction model

ZHANG Yuke

(College of Finance and Statistics, Hunan University, Changsha/Hunan, 410006)

Abstract : Vigorously developing new-energy vehicles is an important measure for our country to promote green development and build a beautiful China. In recent years, the number of new energy vehicles has increased rapidly, but there are many pricing factors in the pricing of new energy vehicle insurance, and the pricing model needs to be improved. This paper innovatively proposes a two-layer model for pricing new energy vehicle insurance based on generalized linear model and XGBoost algorithm. The generalized linear model can be used as the bottom output to ensure the interpretability of the model, and the prediction accuracy of the model can be increased after XGBoost algorithm is added. Therefore, the two-layer prediction model has good performance. Based on a group of actual claims data of new energy vehicles, empirical analysis is carried out. The empirical results show that the prediction effect of the two-layer prediction model is better, the Bayesian information criterion and the model deviation are smaller than the single generalized linear model, and the two-layer model can better predict the extreme loss.

Keywords: New energy vehicle insurance; Generalized linear model; XGBoost algorithm; Two-layer model

作者简介: 张毓科, 男, 湖南大学金融与统计学院应用经济学专业硕士在读研究生, 研究方向为精算理论与实务。