随机森林在新能源车险定价中的应用

卢秋羽

（湖南大学金融与统计学院，湖南长沙，410006）

摘要：当前行业测算的新能源车险基准保费属于车型定价范畴，考虑风险因子不全面，且测算所用广义线性模型仅能识别纯风险损失与定价因子间的线性关系，准确性有待提高。对此，本文在车型定价基础上，将“三电”系统参数、驾驶行为因子、充放电习惯因子、周围环境因子纳入定价因子范围，并使用随机森林算法建立新能源车险定价模型，以提高定价的公平性和科学性。结果表明，品牌、险种决定保费的基准水平，里程、夜间行驶、速度等驾驶行为因子是影响保费波动的重要因素，充电行为因子和周围环境因子是影响保费波动的次重要因素。同时，比起GLM，随机森林有更好的拟合效果和风险区分能力，且随定价因子变化，预测保费与真实损失呈现相同趋势。

关键词：保险学；新能源车险；随机森林

中图分类号：F840.65　　文献标识码：A

1 引言

新能源车是以蓄电池、燃料电池、乙醇等非常规燃料作为动力来源的汽车。根据公安部统计，2022 年全国新能源车保有量约 1310 万辆，占汽车总量的 4.10%；同年全国新注册登记新能源车 535 万辆，同比增长 81.48%，呈高速增长态势。随新能源车普及，新能源车险的规模也不断扩大，2021 年上半年，人保新能源车在保数量 103 万台，保费收入 40.3 亿元，同比增长 60%。然而，相比于传统车险 40%-70%的赔付率，新能源车险的赔付率高达 85%。究其原因，是新能源车相比于传统燃油车有特殊风险，行业应该为新能源车设置专属保险条款并单独进行基准保费测算。2021 年底，《新能源汽车商业保险专属条款（试行）》与《新能源汽车商业保险基准纯风险保费表（试行）》同时发布，标志我国新能源车险迈出新一步。

与传统燃油车相比，新能源车的新增风险体现在两方面。一方面，新能源车险的出险概率增大：①新能源车专属条款将“三电”系统（电池、驱动电机、电动控制器）和车辆的停放、充电状态纳入保障范围；②新能源车的驾驶人多是年轻人，驾驶风险相对较大；③新能源车提速快、失速快，与传统燃油车在驾驶技术上存在差异，从传统燃油车转到新能源车的驾驶人容易出险。另一方面，新能源车险的索赔次均增加：新能源车的电池设计、智能化装置、精密配件及制造工艺等价格整体偏高，零整比也相应偏高。

当前，行业按照被保险人类别、使用性质、座位数、厂牌型号、车龄和区域制定新能源车险分级费率，未考虑电池和驾驶人，仍属于车型定价范畴；同时，测算方法为广义线性模型（GLM），仅限于识别纯风险损失与定价因子间的线性关系。因此，亟待为新能源车险定价模型挖掘更多关于电池和驾驶人的因子，设计更科学合理的风险分级制度。

2 文献综述

目前，关于新能源车险定价的研究集中于定价因子和定价模型。

从定价因子角度，当前行业新能源车险基准保费基于使用性质、座位数、车龄、品牌等测算，仍属于车型定价的范畴。金佳钰（2017）总结出三类车险定价因子：①从车因子，重

点关注零整比和常用配件负担指数、物理碰撞试验结果、汽车自身的设计和结构参数等；②从人因子，重点关注驾驶行为、年龄、驾龄、性别等；③使用状况，重点关注行驶里程、行驶时间段、行驶区域和行驶环境[1]。目前，被广泛用于车险定价的因子有：品牌、车系等从车因子，性别、年龄等从人因子，以及区域等环境因子。近年来，依托数据建设和车联网技术的发展，许多学者挖掘出新的车险定价因子。魏丽等（2018）[2]和 Zhang 等（2022）[3]分别提出将零整比、碰撞测试结果等从车因子加入定价，增加车型定价的维度，然而目前进行零整比测算和碰撞测试的主要是传统燃油车，能获得这部分数据的新能源车的数量极少，故本文暂不考虑这部分因子。Huang 等（2019）提出与车企开展 UBI 合作将夜间行驶、疲劳驾驶等驾驶人行为因子加入定价[4]，从车型定价走向“车+人”定价，由于驾驶人确实是影响车辆是否出险的重要因素，且当前许多新能源车在出厂时已配备驾驶行为数据采集装置，因此定价因子理应包含驾驶行为因子。苏洁（2021）[5]、贾璇等（2022）[6]和傅勇等（2022）[7]指出新能源车涉及电池等核心部件的损毁率较高，且这些核心部件的维修成本较高，最终造成新能源车险的高赔付率，因此定价因子理应包含电池相关参数与充放电行为因子。

从定价模型角度，当前行业使用的测算模型主要是 GLM。由于 GLM 只能识别定价因子与纯风险损失的线性关系，风险分级效果有限，学者们尝试将机器学习算法用于识别定价因子与纯风险损失的非线性关系，以获得更好的风险分级效果。如 Yunos 等（2016）尝试将人工神经网络（ANN）用于车险定价[8]；孟生旺等（2017）对比 GLM、支持向量机（SVM）和多层感知器（MLP）对索赔发生概率的预测效果，发现 MLP 的预测精度稍优于 GLM，但结构复杂得多，而 SVM 无论从预测精度还是效率上都比不过 GLM[9]；张碧怡等（2019）发现随机森林（RF）和 XGBoost 能提高车险索赔频率预测精度，而且特征重要性能与 GLM 保持一致[10]。

当前，国内对新能源车险定价的实证分析文献极少，仅蒋涵（2020）以续航里程和驱动方式等变量为定价因子，构建新能源车险的索赔频率和索赔强度的 GLM 模型[11]。为实现新能源车险的精细化定价，本文在车型定价的基础上，将“三电”系统参数、驾驶行为因子、充放电习惯因子、周围环境因子纳入定价因子范围，使得风险分级的维度从“车辆”扩展至“车辆+驾驶人”；同时，运用随机森林算法建立新能源车险定价模型，以识别 GLM 所不能识别的非线性的风险分级规则，大幅提高定价的公平性和科学性。

3 理论基础

3.1 决策树

决策树（DT）是一种常用的机器学习算法，其实质上是由若干条判断规则组成的树形规则集合，已被广泛应用于回归（Regression）和分类（Classification）问题。由于本文的因变量为车均损失，属于连续型变量，本文将从回归树的角度介绍 DT 的原理。

回归树的生成是一个递归过程，初始时所有样本都在根结点，而后通过不断分裂结点，直至所有样本都落在叶结点中。分裂结点时，将遍历所有因子的所有取值，选择使分裂后组内离差平方和最小的因子及对应分裂阈值。当结点满足下列三个条件之一时，结点被标记为叶结点，否则标记为内部结点。

①结点中样本个数少于最小样本阈值；

②结点中所有样本在所有因子上都取值相同；

③回归树已生长到最大深度。

显然，叶结点没有子结点，内部结点必有子结点。如下图是回归树的结构示例，方框表示叶结点，圆圈表示内部结点。叶结点的权值是该结点中所有样本的因变量取值的均值，作为训练集或测试集所有落在该节点中的样本的预测值。

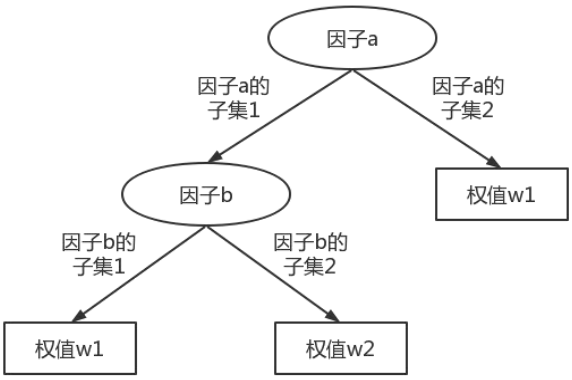

图 1　回归树的结构示例

3.2　随机森林

随机森林（RF）是决策树和 Bagging 的结合。Bagging 是一种并行式集成学习方法，参考自助采样法（Bootstrap Sampling）的思想，对原始数据集做多次随机放回抽样，然后基于每个采样数据集训练基学习器，最后将这些基学习器的预测值结合起来。确定最终预测值时，分类问题一般采用简单投票法，回归问题一般采用简单平均法。实践证明，Bagging 可以减小基学习器的预测方差，提高模型的稳定性。RF 的本质是基学习器为决策树的 Bagging，并进一步引入对自变量集合的抽样。即，RF 同时训练多棵树，每棵树的训练数据集和训练自变量集都不同。比起单棵回归树，RF 通过样本干扰，减少模型对特定样本的依赖，降低过拟合风险；同时，确定最终预测值时结合多棵树的预测结果，提高预测值的稳定性。比起 Bagging，RF 增加对自变量的干扰，能减少模型对特定自变量的依赖，提高其泛化能力。

4　基于随机森林的新能源车险定价模型

4.1　数据来源

本文数据是国内某保险公司 2019 至 2021 年的新能源家用车商业保险的承保理赔数据，经过异常值与缺失值处理、损失和保费的发展、特征工程、连续变量分组处理、根据相关性筛选定价因子等预处理后，剩余 93332 条保单记录，37 个变量，详见下表。

表 1　定价因子

一级分类	二级分类	三级分类	类型
从人因子	使用程度	日均行程数（单位：次）	连续
从人因子	使用程度	单次行程平均行驶里程（单位：0.1km）	连续
从人因子	使用程度	平均速度（单位：km/h）	连续
从人因子	夜间行驶	夜间（20:00-4:00）行程数占比（单位：%）	连续
从人因子	夜间行驶	深夜（00:00-4:00）行程数占比（单位：%）	连续
从人因子	夜间行驶	深夜（00:00-4:00）单次行程平均行驶里程（单位：0.1km）	连续
从人因子	夜间行驶	深夜（00:00-4:00）行驶平均速度（单位：km/h）	连续
从人因子	早晚高峰行驶	早晚高峰（7:00-9:00/17:00-19:00）行程数占比（单位：%）	连续
从人因子	节假日行驶	节假日行程数占比（单位：%）	连续
从人因子	疲劳驾驶	疲劳驾驶行程数占比（单位：%）	连续

（续表）

从人因子	疲劳驾驶	严重疲劳驾驶行程数占比（单位：%）	连续
从人因子	疲劳驾驶	疲劳驾驶单次行程平均行驶时长（单位：s）	连续
从人因子	疲劳驾驶	严重疲劳驾驶单次行程平均行驶时长（单位：s）	连续
从人因子	短行程行驶	短行程数占比（单位：%）	连续
从人因子	短行程行驶	单次短行程平均行驶时长（单位：s）	连续
从人因子	高速行驶	高速行驶时长占比（单位：%）	连续
从人因子	高/低温行驶	行车平均最高温度（单位：℃）	连续
从人因子	高/低温行驶	高温行驶时长占比（单位：%）	连续
从人因子	高/低温行驶	低温行驶时长占比（单位：%）	连续
从人因子	充电习惯	日均充电次数（单位：次）	连续
从人因子	充电习惯	单次充电平均时长（单位：s）	连续
从人因子	充电习惯	充电起始剩余电量（单位：%）	连续
从人因子	充电习惯	充电结束剩余电量（单位：%）	连续
从人因子	充电习惯	充电平均最高温度（单位：℃）	连续
从人因子	充电习惯	快充次数占比（单位：%）	连续
从人因子	充电习惯	高温快充次数占比（单位：%）	连续
从人因子	充电习惯	高温充电时长占比（单位：%）	连续
从人因子	充电习惯	低温充电时长占比（单位：%）	连续
从人因子	用电习惯	百公里耗电量（单位：kwh/100km）	连续
从人因子	用电习惯	行车起始剩余电量（单位：%）	连续
从人因子	用电习惯	行车结束剩余电量（单位：%）	连续
从车因子		车龄（单位：年）	连续
从车因子		品牌	离散
从车因子		动力类型（水平：电动/插电式混合动力）	离散
从车因子		电池类型（水平：三元锂电池/磷酸铁锂电池）	离散
从车因子		电池包总容量（单位：ah）	连续
险种		险种（水平：主全/交三/单交）	离散

4.2 随机森林车险定价模型

经过调参操作后，确定 RF 纯风险保费预测模型的参数如下：

表 2 随机森林的参数

参数	参数取值
n_estimators	100
max_features	0.6
max_samples	0.7
max_depth	9
min_samples_leaf	800
因变量	车均损失
自变量是否标准化	是
自变量数量	37
数据量	93332

基于预处理后的数据集，设置随机数种子为 5，按照上表参数，建立 RF 纯风险保费预测模型。RF 预测纯风险保费的分布如下图，多集中于 1000 至 4000；也存在超过 10000 的保费，对应样本的单次行程平均行驶里程均超过 36.8km。

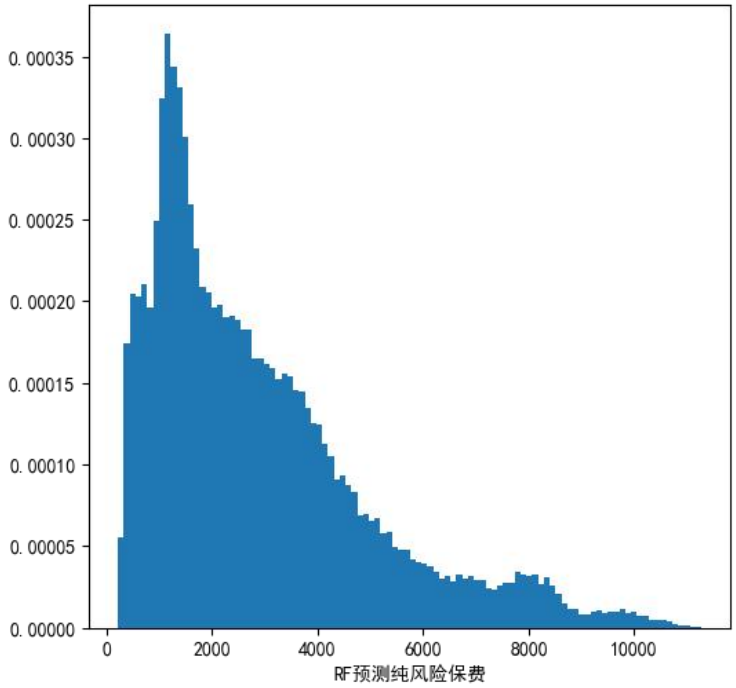

图 2　随机森林预测纯风险保费的直方图

4.3　模型效果评价

为评价随机森林纯风险保费预测模型的效果，本文设置了两个对照组：

①基准保费：由原始数据集提供；

②GLM 预测纯风险保费：基于预处理后的数据集建立 GLM 得到；

并从均方误差、提升曲线和一致性曲线三个角度对比随机森林预测纯风险保费与两个对照组的表现。

4.3.1　均方误差

均方误差（RMSE）描述的是预测纯风险保费与真实损失间的离差，RMSE 越小，说明模型拟合程度越好。由下表，RF 的拟合效果最好、GLM 次之、基准保费最差。

表 3　均方误差

模型	RMSE
基准保费	20559
GLM 预测纯风险保费	20492
RF 预测纯风险保费	20476

4.3.2　提升曲线

提升曲线用于分析真实损失是否随风险级别的提高而增加。绘制提升曲线的步骤如下：①根据预测纯风险保费将样本划分为从低到高的 10 个风险等级；②计算每个风险等级内样本的平均真实损失；③将每个风险等级及其对应的平均真实损失绘于图上。若保费设置合理，则提升曲线是一条平滑上升的曲线。

由下图，基础保费的提升曲线整体呈现缓慢上升趋势；GLM 的提升曲线整体呈现上升趋势，但在第 7 组出现轻微的保费倒挂；RF 的提升曲线整体呈现平滑上升趋势，未出险保费倒挂的情况。综合来看，RF 是最具合理性、公平性和稳定性的模型；GLM 能区分不同风险等级的样本，但可能存在保费倒挂的情况；基准保费区分不同风险等级样本的能力较弱。

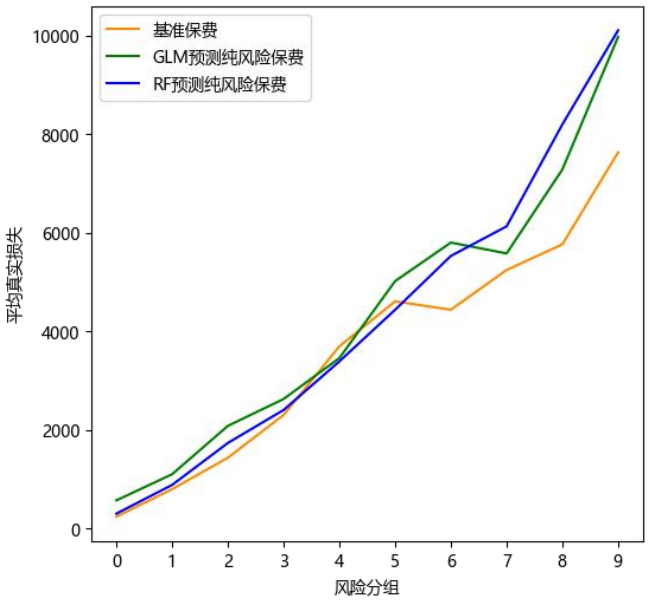

图 3　提升曲线

4.3.3　一致性曲线

一致性曲线用于分析，随单因子的变化，平均预测纯风险保费和平均真实损失是否呈现相同的趋势。这类似于单因素分析，但两者目的不同：单因素分析的目的是在数据筛选阶段挑选与平均损失相关的因子；而一致性曲线的目的是检验预测纯风险保费与单因子的关系是否同实际保持一致。本文将从以下个角度，分析一致性曲线。

（1）快充次数占比

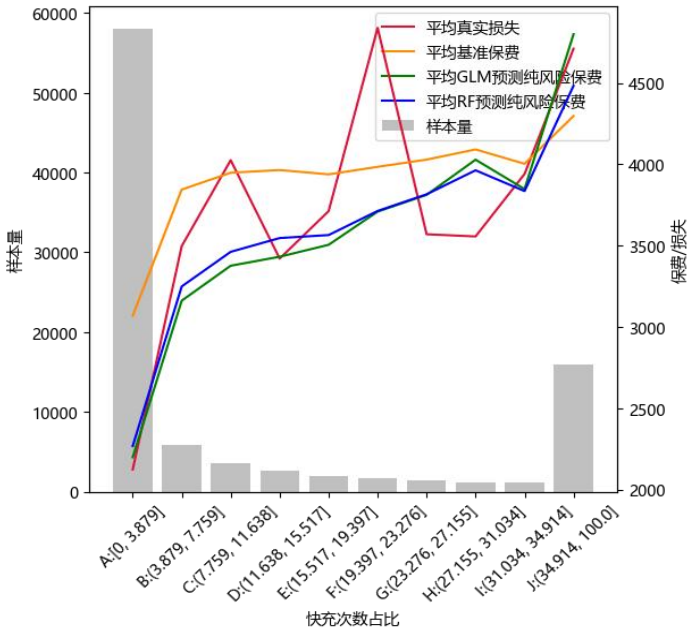

图 4　基于快充次数占比的一致性曲线

由上图，在快充次数占比的分组上，基准保费、GLM、RF 均和平均真实损失保持同一趋势。随快充次数占比增大，保费呈现增大趋势。究其原因，是快充通过提高电压或电流来提高充电功率，以达到快速充电的目的；然而，电流过大时，电池负极表面的半透膜容易出现轻微破裂，电解液也容易分解，导致电池不可逆的损耗。因此，快速充电是以牺牲电池循环寿命为代价的，为延长电池使用寿命，应避免频繁快充。

（2）充电起始剩余电量和充电结束剩余电量

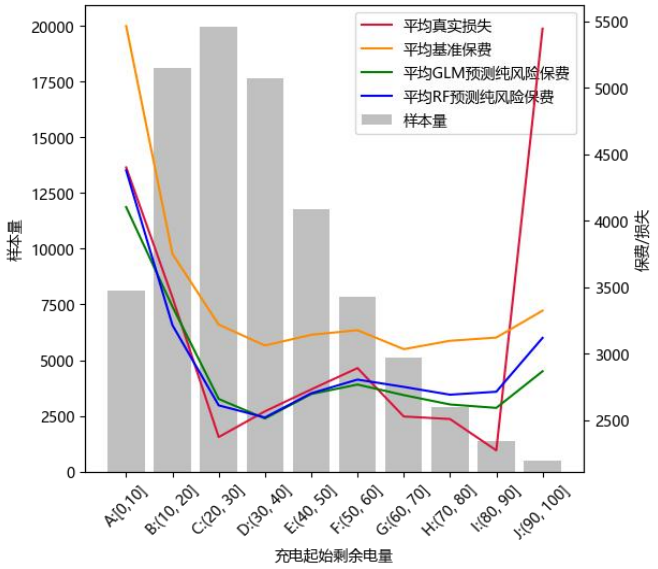

图 5　基于充电起始剩余电量的一致性曲线

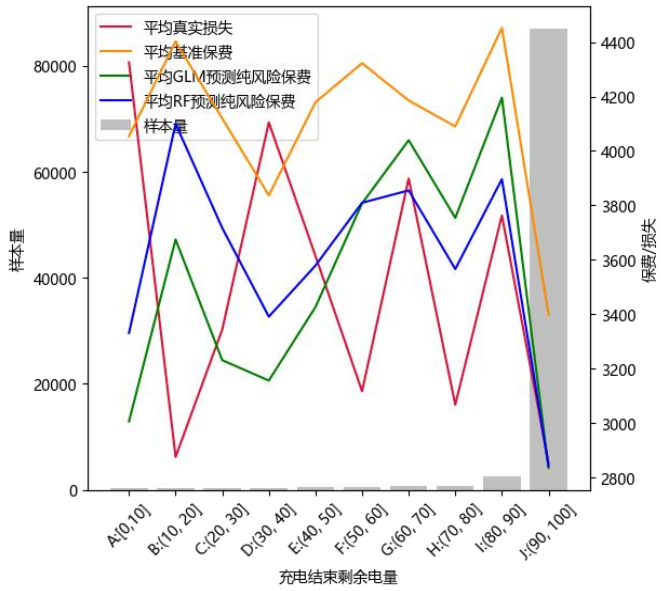

图 6　基于充电结束剩余电量的一致性曲线

在充电起始剩余电量的分组上，基准保费、GLM、RF 均和平均真实损失保持同一趋势，但在在充电结束剩余电量的分组上不能。究其原因，是大部分车主在充电时都会选择充满，导致充电结束剩余电量高度集中于 90%-100%，而其他区间的样本量很少，组内平均真实损失失去统计意义。

在充电起始剩余电量少于 20%或大于 90%时，保费较高。究其原因，当电池起始剩余电量少于 20%时，处于过度放电状态，电池中碳层容易发生塌陷，导致充电时锂离子无法进入碳层，从而影响电池的实际容量；当充电起始剩余电量大于 90%时，处于过度充电状态，碳层中容易过多嵌入锂离子，导致放电时部分锂离子无法脱嵌，影响电池使用性能。

（3）百公里耗电量

由下图，在百公里耗电量的分组上，基准保费、GLM、RF 均和平均真实损失保持同一趋势。随百公里耗电量增大，保费呈现先减小后增大的趋势。一方面，保费在 A、B 两组水平较高的原因是：两组中 95%的动力类型都是插电式混合动力，相比纯电动具有更复杂的结

构，对应更高的车辆价格和保费。另一方面，保费后增大的原因是：百公里耗电量越大，一般对应车辆价格越高或驾驶人有更多危险行为，从而保费越高。

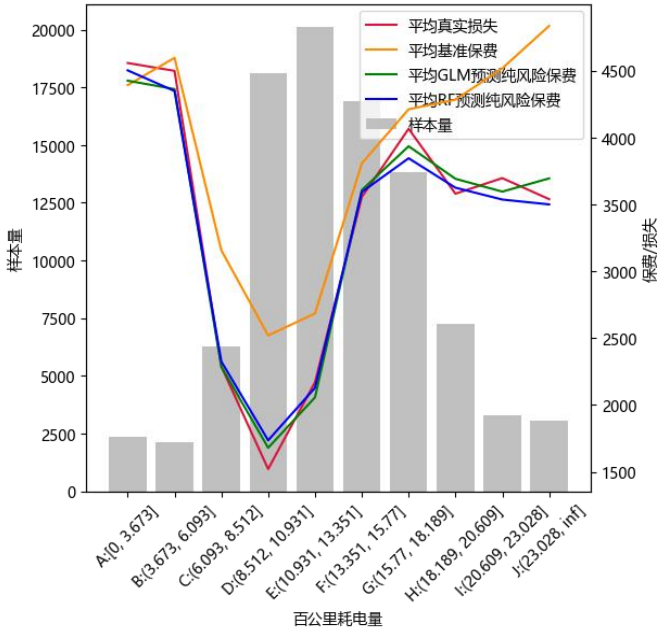

图 7　基于百公里耗电量的一致性曲线

4.4　特征重要性

RF 有提供特征重要性来度量不同因子对预测纯风险保费的影响程度。因子 j 的特征重要性的计算过程是：第一步，在随机森林的所有子树中，计算因子 j 被用于分裂节点所带来均方误差的减小程度的总和；第二步，将所有因子的特征重要性进行归一化，即为因子 j 的特征重要性。

由下图，①品牌是最重要的因子，一般整车价格是决定保费的主要原因，当两辆车的整车价格相当时，零整比越高的品牌对应保费越高；②单次行程平均行驶里程是次重要因子，长途行驶条件下，驾驶人注意力下降，出险概率增加，同时容易导致电池和电机的损害；③夜间行驶系列因子（深夜行程数占比、深夜行驶平均速度、深夜单次行程平均行驶里程）是影响赔付的重要因素，因为夜间照明条件下，驾驶人对路况的判断能力下降，出险概率增加。

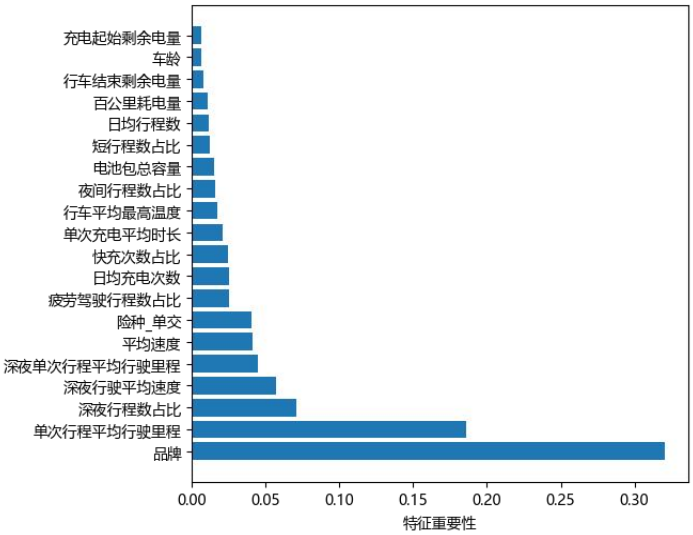

图 8　RF 的特征重要性

综上所述，品牌决定保费的基准线，里程、夜间行驶、速度等驾驶行为因子是影响保费的重要因素，充电习惯因子和周围环境因子是影响保费的次重要因素。

最后，将 RF 的特征重要性与 GLM 的斜率系数对比。一般，用 GLM 的斜率系数表示因子对预测纯风险保费的影响，其符号代表影响方向、数值代表影响程度。给定显著性水平 10%，GLM 新能源车险定价模型的显著变量系数如下：

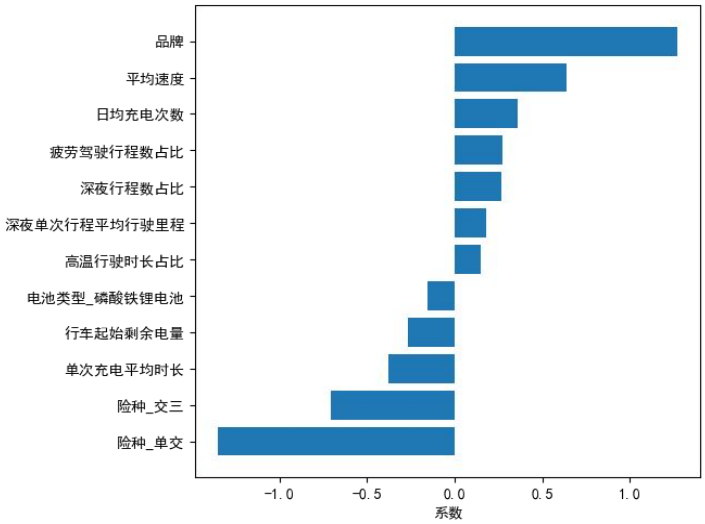

图 9　GLM 纯风险保费预测模型的显著变量系数

由上图，①品牌是最重要的因子；②单次行程平均行驶里程并非次重要因子，取而代之的是险种；③夜间行驶系列因子的重要性降低，取而代之的是充电习惯因子。这是因为 GLM 仅能识别线性关系，更重视与平均真实损失线性相关性较强的因子，而 RF 能同时识别线性与非线性关系。

5　结论

本文在车型定价的基础上，将“三电”系统参数、驾驶行为因子、充放电习惯因子、周围环境因子纳入定价因子范围，丰富定价因子体系；同时，运用随机森林算法建立新能源车险定价模型，以识别 GLM 所不能识别的非线性的风险分级规则，大幅提高定价的公平性和科学性。有以下结论：（1）关于定价因子：品牌、险种决定保费的基准水平，里程、夜间行驶、速度等驾驶行为因子是影响保费波动的重要因素，充电习惯因子和周围环境因子是影响保费波动的次重要因素。（2）关于定价模型：比起 GLM，随机森林有更好的拟合程度和风险区分能力，且随定价因子变化，预测保费与真实损失呈现相同趋势。经验证，随机森林的特征重要性能与 GLM 斜率系数大体上保持一致，但存在一定差异；究其原因，是 GLM 仅能识别线性规则，而随机森林能识别非线性规则。

综上所述，本文提出以下建议：

（1）保险公司积极与车企展开 UBI 合作：越来越多新能源车在出厂时已配置驾驶行为数据收集装备，由车企负责加工处理与管理；同时，高级驾驶辅助系统的出现使得车企收集的数据不局限于里程、速度等传统驾驶行为因子，还包括路况预警等信息，能更全面地描述标的车辆的风险。保险公司可积极与车企展开 UBI 合作，通过分析对应品牌的保费规模、赔付风险、因子提升效果等，决定是否有合作的必要。

（2）保险公司可积极尝试将机器学习算法引入车险定价，以辅助 GLM，提高定价的合理性与公平性。

参考文献

【1】金佳钰. 商业车险个性化风险分析[J]. 汽车工业研究,2017(12):42-45.

【2】魏丽,杨斐滟. 我国商业车险改革评析[J]. 保险研究,2018(5):16-32.

【3】Lin Zhang, JiangPing Yang. Research on auto insurance pricing based on LightGBM[C] //International Conference on Cyber Security, Artificial Intelligence, and Digital Economy (CSAIDE 2022). Society of Photo-Optical Instrumentation Engineers(SPIE), 2022, 12330: 123300D.

【4】Yifan Huang, Shengwang Meng. Automobile insurance classification ratemaking based on telematics driving data[J]. Decision Support Systems,2019,127:113156.

【5】苏洁. 新能源车险为何定价难[N]. 中国银行保险报,2021-09-29(005).

【6】贾璇. 新能源汽车进入专属保险时代[J]. 中国经济周刊,2022(1):74-75.

【7】傅勇,袁小康. 新能源车求解零整比难题[N]. 经济参考报,2022-12-30(006).

【8】Zuriahati Mohd Yunos, Aida Ali, Siti Mariyam Shamsyuddin, Noriszura Ismail, Roselina Sallehuddin. Predictive Modelling for Motor Insurance Claims Using Artificial Neural Networks[J]. International Journal of Advances in Soft Computing and its Applications, 2016, 8(3): 160-172.

【9】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.

【10】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.

【11】蒋涵. 基于广义线性模型的新能源汽车保险费率厘定研究[D].山东大学,2020.

New Energy Vehicle Insurance Pricing Based on Random Forest

LU Qiuyu

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

Abstract： The base premium of new energy vehicle insurance which is widely used by insurance companies was rated considering brand only and using GLM which can identify the linear relationship between loss and pricing factor only. To improve the fairness of premium, this thesis consider pricing factors about environment, electricity, drivers, charging and discharging habits, and construct a new energy vehicle insurance pricing model based on random forest. The results show that brand and type of insurance determine the level of premium, factors about drivers such as mileage, night driving, and speed are important factors which affect premium, and factors about charging habits are secondary important. Moreover, compared with GLM, random forest fits better and has larger premium discrimination between clients of low and high risk, and the prediction shows the same trend as the loss as a factor changes.

Keywords: Insurance; New energy vehicle insurance; Random forest