Statistics and Application 统计学与应用, 2026, 15(3), 55-65
Published Online March 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.153055

Hans汉斯

基于贝叶斯加性回归树的风险保费建模方法研究

王艺霖

河北工业大学理学院，天津

收稿日期：2026年2月10日；录用日期：2026年3月3日；发布日期：2026年3月12日

摘　要

车险索赔数据普遍存在索赔发生不均衡、索赔金额重尾分布及风险特征高度非线性等问题，给风险保费建模带来挑战。本文提出一种融合机器学习与期望分位数(Expectile)保费原理的风险保费建模框架。该框架由“两阶段纯保费模型”和“贝叶斯期望分位数加性回归树(BEART)风险附加模型”两部分构成，分别用于刻画车险的出险概率和条件累计索赔额情况及索赔分布的尾部风险特征。实证结果表明，该框架在风险识别、尾部风险刻画及预测稳定性方面具有良好表现，可为复杂车险风险环境下风险保费厘定提供可靠的建模依据。

关键词

风险保费，贝叶斯加性回归树，期望分位数，非对称平方损失函数，评价指标

Research on Risk Premium Modeling Methods Based on Bayesian Additive Regression Trees

Yilin Wang

School of Sciences, Hebei University of Technology, Tianjin

Received: February 10, 2026; accepted: March 3, 2026; published: March 12, 2026

Abstract

Auto insurance claims data are commonly characterized by imbalanced claim occurrence, heavy-tailed claim severity distributions, and highly nonlinear risk features, posing substantial challenges for risk premium modeling. This paper proposes a risk premium modeling framework that integrates

文章引用: 王艺霖. 基于贝叶斯加性回归树的风险保费建模方法研究[J]. 统计学与应用, 2026, 15(3): 55-65.
DOI: 10.12677/sa.2026.153055

machine learning techniques with the Expectile premium principle. The framework consists of two components: a two-stage pure premium model and a Bayesian Expectile Additive Regression Tree (BEART)-based risk loading model, which are employed to characterize claim occurrence probability, conditional cumulative claim amounts, and tail risk features of the claims distribution, respectively. Empirical results demonstrate that the proposed framework exhibits strong performance in risk discrimination, tail risk characterization, and predictive stability, providing a reliable modeling basis for risk premium determination in complex motor insurance risk environments.

Keywords

Risk Premium, Bayesian Additive Regression Trees, Expectile, Asymmetric Squared Loss Function, Evaluation Metrics

1. 引言

机动车辆保险因索赔频率高、赔付呈右偏厚尾且数据不平衡，一直是精算研究重点。随费率市场化推进，提升风险识别与定价科学性成为核心。杨亮和孟生旺[1]指出风险保费由纯保费与风险附加构成，前者补偿期望赔付，后者覆盖波动与尾部不利偏差。传统车险费率厘定研究多基于广义线性模型对纯保费进行建模，并在此基础上采用期望值或标准差保费原理引入风险附加。然而，该类方法依赖人为设定的风险附加系数，主观性较强，且以均值或方差为核心，难以刻画赔款分布的非对称性与厚尾特征，对极端损失风险反映不足。

近年来，分位数回归和 Expectile 回归方法在金融风险度量领域受到了广泛关注，有学者将其引入来进行车险定价的探索研究，Heras 等[2]和 Yang 等[3]分别提出分位数保费原理和期望分位数保费原理，通过将风险保费作为整体进行建模，在控制赔款超过风险保费概率的同时引入不对称损失权重，从而兼顾统计效率与尾部风险刻画能力，为车险高风险定价提供了更具针对性的建模框架。后续部分学者进一步将分位数回归和 Expectile 回归结合随机森林、XGBoost 和神经网络等机器学习方法以提升预测能力[4]-[6]，这些研究多停留在模型的理论层面，且缺乏基于复杂车险数据的实证探索。Chipman 等[7]提出贝叶斯加性回归树兼具非参数建模与不确定性量化优势，但在车险领域中尚未充分应用。因此，本文融合贝叶斯加性回归树与 Expectile 思想，构建风险保费建模框架，以系统刻画纯保费与风险附加，为复杂风险环境下的精算定价提供新思路。

2. 模型构建与技术路线

2.1. 研究设计

Heras 等[2]提出分位数保费原理(Quantile premium principle, QPP)，表达式为：

$$H\left(Y_i \mid \boldsymbol{x}_i\right)=\mathbb{E}\left(Y_i \mid \boldsymbol{x}_i\right)+\varphi\left[Q_{Y_i}\left(\tau \mid \boldsymbol{x}_i\right)-\mathbb{E}\left(Y_i \mid \boldsymbol{x}_i\right)\right] \tag{1}$$

Yang 等[3]提出 Expectile 保费原理(Expectile premium principle, EPP)，表达式为：

$$H\left(Y_i \mid \boldsymbol{x}_i\right)=\mathbb{E}\left(Y_i \mid \boldsymbol{x}_i\right)+\varphi\left[expectile_{Y_i}\left(\tau \mid \boldsymbol{x}_i\right)-\mathbb{E}\left(Y_i \mid \boldsymbol{x}_i\right)\right] \tag{2}$$

EPP 基于 Expectile 回归计算出来的风险保费作为一种新颖的风险测度，满足一致性风险测度的四条性质要求即平移不变性、单调性、正齐次性、次可加性，且对极端值更敏感，因此可以更好地刻画总索赔额条件分布的尾部特征。郭哲琦和高苏浩[8]提出两阶段 Expectile 回归来预测每份保单的风险保费使得计算结果更加合乎定价逻辑。

因此，本文参考 Expectile 保费原理及两阶段建模思想，构建车险风险保费建模框架。首先，构建基于机器学习的两阶段纯保费模型，分别建立出险概率模型与在损失发生条件下的累计索赔额预测模型从而获得纯保费估计；其次，在风险附加环节，创新性地将贝叶斯加性回归树引入车险定价研究，并与期望分位数思想相结合，构建贝叶斯期望分位数加性回归树模型，通过给定分位数水平 τ 估计条件期望分位数以刻画尾部风险，最终按照式(2)计算得到完整风险保费，其中 φ 为风险厌恶系数，可结合实际情况人为给定。整体流程如图 1 所示。

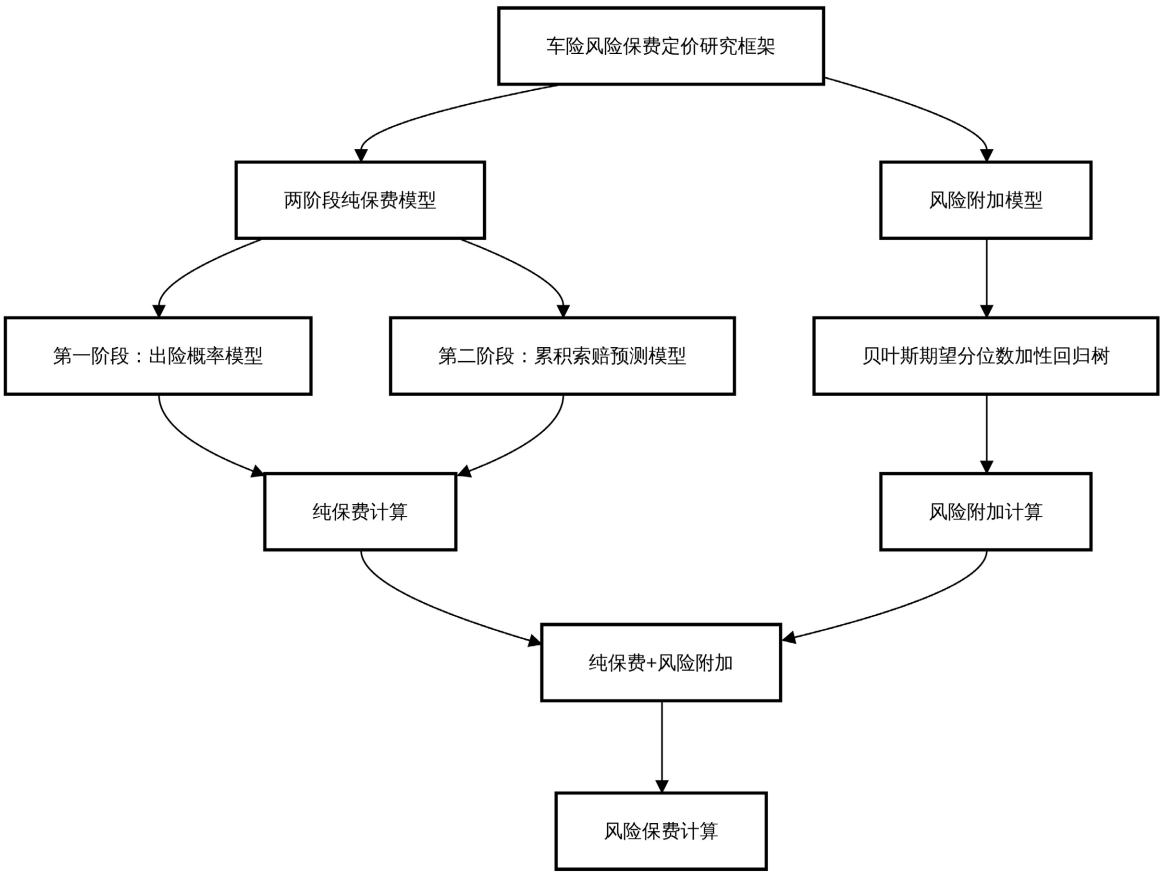

Figure 1. Research framework of the proposed methodology
图 1. 论文研究框架流程图

2.2. 模型评估指标说明

2.2.1. 分类模型评估指标

在分类模型评估中，引入 AUC、AUPRC、Accuracy、Precision、Recall 和 F1 等多种评价指标，从整体判别能力、少数类识别效果以及预测结果的准确性与稳健性等不同维度综合衡量模型性能，避免单一指标在类别不平衡或应用约束条件下产生偏误，从而更全面、客观地评价分类模型的实际适用性。其中，Accuracy、Precision、Recall 和 F1 都是基于真实标签与预测标签构建的混淆矩阵进行计算，评价指标的

表达式分别如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

2.2.2. 回归模型评估指标

在回归模型评估中，MSE、RMSE 和 MAE 等指标用于从整体误差幅度、误差尺度及稳健性角度衡量模型对响应变量条件均值的拟合精度；而在分位数回归与期望分位数回归模型中，引入 WMAD 和 WMSE 等加权误差指标，通过强调不同样本，尤其是尾部观测值的重要性，能够更有针对性地评估模型对条件分布非对称性及尾部风险特征的刻画能力，从而增强模型评价结果与风险度量目标之间的一致性。假设 y_i 为第 i 份保单实际损失的观测值，$\hat{f}(x_i)$ 为对应的预测值，样本数为 n，评价指标的表达式分别如下：

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2} \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{f}(x_i)\right| \tag{9}$$

$$\text{WMAD} = \frac{\sum_{i=1}^{n} w_i \left|y_i - \hat{f}(x_i)\right|}{\sum_{i=1}^{n} w_i} \tag{10}$$

$$\text{WMSE} = \frac{\sum_{i=1}^{n} w_i \left(y_i - \hat{f}(x_i)\right)^2}{\sum_{i=1}^{n} w_i} \tag{11}$$

3. 理论模型

3.1. 预备知识

3.1.1. 贝叶斯加性回归树(BART)

Chipman 等[7]提出的贝叶斯加性回归树可以看作是梯度提升方法的贝叶斯版本。Boosting 通过迭代地将连续残差与协变量拟合来构造其组成部分。相比之下，BART 通过使用正则化来创建一个贝叶斯“树和”模型，然后将每棵树约束为弱学习者。BART 使用 MCMC 算法，通过不断生成和更新回归树的结构、每个树的叶子节点参数以及整体的误差方差，模型预测值是所有树预测值的总和。

3.1.2. Expectile 回归

Newey 和 Powell [9]提出线性 Expectile 回归模型，原理是通过最小化非对称平方损失函数进行建模，并证明了其估计量服从渐近正态分布。Expectile 回归不仅兼具均值回归和分位数回归的优点，而且有计算简便，预测精度高等优良特性。特别的是，线性 Expectile 回归对异常值较为敏感，这一性质使其在金融风险测度领域得到广泛应用。

期望分位数定义为：

$$Q(\tau)=\arg\min_{\theta} E\left\{\left|\tau-I\left(\left[Y_i-\theta\right]<0\right)\right|\left[Y_i-\theta\right]^2\right\} \tag{12}$$

Expectile 回归是分位数回归的“平方损失版本”，其目标是最小化非对称平方损失函数：

$$L_{\tau}(y,\hat{y})=\begin{cases}\tau\cdot(y-\hat{y})^2, & \text{if } y\geq\hat{y}\\ (1-\tau)\cdot(y-\hat{y})^2, & \text{if } y<\hat{y}\end{cases} \tag{13}$$

其中，$\tau\in(0,1)$ 是目标 Expectile 水平。

3.2. 贝叶斯期望分位数加性回归树(BEART)

曹桃云和张日权[10]指出标准 BART 通常假设误差项服从正态分布，其对应的似然函数等价于平方损失，因此模型的后验推断本质上针对条件均值展开，难以直接刻画期望分位数回归所需的非对称平方损失。金娇等[11]提出为在贝叶斯框架下实现 Expectile 估计，需要对模型的误差分布假设进行扩展，引入一种与 Expectile 损失函数严格对应的概率模型。

因此，在 BEART 模型中，假设响应变量满足如下模型结构：

$$y_i=f(x_i)+\epsilon_i \tag{14}$$

其中，$f(x_i)=\sum_{j=1}^{m}g_j(x_i;T_j,M_j)$，$g_j$ 为第 j 棵回归树；误差项 $\epsilon_i\sim\mathrm{AND}(0,\sigma^2,\tau)$。

对应于 Expectile 损失函数的非对称正态分布，其密度函数为概率分布，通过对全空间积分并推导归一化因子，最终得到非对称正态分布 AND 的概率密度函数为：

$$p(y\mid\mu,\sigma,\tau)=\frac{2}{\sqrt{\pi}\sigma}\cdot\frac{\sqrt{\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}}\cdot\exp\left(-\frac{1}{\sigma^2}\begin{cases}\tau(y-\mu)^2, & y\geq\mu\\ (1-\tau)(y-\mu)^2, & y<\mu\end{cases}\right) \tag{15}$$

其中，μ 是位置参数，σ 是尺度参数，τ 控制分布的非对称性。

式(15)所对应非对称正态分布 AND 的负对数似然函数为：

$$NLL=\sum_{i=1}^{n}\left[\tau(y_i-\mu)^2\cdot I(y_i\geq\mu)+(1-\tau)(y_i-\mu)^2\cdot I(y_i<\mu)\right] \tag{16}$$

通过以上推导可以看到，AND 的负对数似然函数与 Expectile 回归的目标损失函数在形式上完全一致，因此最大化 AND 的似然等价于最小化 Expectile 损失函数。通过将 AND 作为误差分布引入 BART 模型，即可在贝叶斯框架下实现从对称平方损失到 Expectile 非对称平方损失的转换。

3.2.1. 联合概率分布分解

在 BEART 模型中，回归函数由多棵回归树的加性结构组成，表达式为：

$$f(x_i)=\sum_{j=1}^{m}g_j(x_i;T_j,M_j) \tag{17}$$

其中，g_j 为第 j 棵回归树，T_j 为树结构，M_j 为对应的叶节点参数集合。

在给定模型假设与先验分布的条件下，BEART 的联合后验分布可分解为似然项与先验项的乘积形式，表达式为：

$$\underbrace{p\left(\{T_j,M_j\},\sigma^2,\tau\mid y_i\right)}_{\text{后验}}=\underbrace{\prod_{i=1}^{n}p\left(y_i\mid\{T_j,M_j\},\sigma^2,\tau\right)}_{\text{似然项}}\cdot\underbrace{\prod_{j=1}^{m}p(T_j)\,p(M_j\mid T_j)}_{\text{树结构和叶节点先验}}\cdot\underbrace{p(\sigma^2)}_{\text{方差先验}} \tag{18}$$

其中，似然项由 AND 诱导，先验项分别对应树结构、叶节点参数及误差方差。

3.2.2. 似然函数

在贝叶斯框架下，似然函数通过假设观测值在给定模型参数条件下相互独立，刻画数据生成机制与模型参数之间的关系。BEART 模型采用非对称正态分布 AND 作为误差分布，以构建与 Expectile 损失函数严格对应的似然结构。该分布通过对正负残差施加不同权重，使模型能够直接刻画条件 Expectile，而非条件均值。BEART 的似然函数表达式为：

$$\begin{aligned} L\left(\boldsymbol{y} \mid T_j, M_j, \sigma^2\right) &= \prod_{i=1}^{n} p\left(y_i \mid \left\{T_j, M_j\right\}, \sigma^2, \tau\right) \\ &= \prod_{i=1}^{n} \frac{2}{\sqrt{\pi}\sigma} \cdot \frac{\sqrt{\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}} \cdot \exp\left(-\frac{1}{\sigma^2} \begin{cases} \tau\left(y - f\left(x_i\right)\right)^2, & y \geq f\left(x_i\right) \\ (1-\tau)\left(y - f\left(x_i\right)\right)^2, & y < f\left(x_i\right) \end{cases} \right) \end{aligned} \tag{19}$$

为便于在贝叶斯推断过程中处理非对称性，引入权重变量 w_i，用于刻画残差方向对似然的影响，权重表达式为：

$$w_i = \begin{cases} \tau, & \text{if } \varepsilon_i \geq 0 \\ 1-\tau, & \text{if } \varepsilon_i < 0 \end{cases} \tag{20}$$

此时，权重由残差符号决定，并在每次 MCMC 迭代中根据当前模型拟合结果动态更新。借助该权重表示，非对称似然可等价转化为加权正态似然形式，从而在后验采样过程中统一用于树结构更新、叶节点参数更新以及方差参数更新，此时非对称正态分布可等价表示为：

$$p\left(y_i \mid f\left(x_i\right), v_i, \sigma\right) \propto \exp\left(-\frac{w_i}{2\sigma^2}\left(y_i - f\left(x_i\right)\right)^2 \right) \tag{21}$$

3.2.3. 先验分布

为在引入 Expectile 非对称性的同时保持模型的正则化特性与计算稳定性，BEART 在先验层面整体继承标准 BART 设定，仅通过似然函数刻画非对称结构。

(1) 树结构先验

通过对每棵回归树的结构 T_j，施加正则化先验，以限制树的复杂度并防止过拟合。树的生长过程通过深度相关的概率控制，较深节点被分裂的概率逐渐减小，从而鼓励生成浅层回归树。分裂变量与分裂点在给定候选集合内均匀选取。

(2) 叶节点参数先验

给定树结构 T_j，叶节点参数 $M_j = \left\{\mu_{jk}\right\}$ 服从独立正态先验，超参数 σ_μ 控制单棵树对预测的贡献强度，超参数设定与标准 BART 保持一致，确保回归函数由多棵弱树平滑叠加得到。

$$\mu_{jk} \sim N\left(0, \sigma_\mu^2\right) \tag{22}$$

(3) 方差先验

对误差项的方差 σ^2 施加逆伽马先验以保持与加权正态似然的共轭性，便于后续 Gibbs 采样更新。推荐 $a = 3, b = 1$。

$$\sigma^2 \sim \text{Inverse-Gamma}\left(a, b\right) \tag{23}$$

3.2.4. 后验更新——MCMC 算法设计修正

在上述先验设定与 AND 似然结构下，BEART 的后验推断可通过 MCMC 方法实现。更新流程在 BART 框架下展开，并在似然项中体现 Expectile 非对称性。

(1) 更新树结构 T_j

树结构的变化如分裂或剪枝会涉及离散参数，无法直接采样，选择通过 Metropolis-Hastings 算法进行更新，其接受概率由加权似然比与先验比共同决定。由于似然项中引入了权重 w_i，树结构的优劣由加权残差平方和决定，从而使树划分更倾向于刻画目标 Expectile 水平。具体计算如下：

接受概率：

$$\alpha = \min\left(1, \frac{p\left(\boldsymbol{y} \mid T_j^*, M_j^*, \sigma^2\right) p\left(T_j^*\right) q\left(T_j \mid T_j^*\right)}{p\left(\boldsymbol{y} \mid T_j, M_j, \sigma^2\right) p\left(T_j\right) q\left(T_j^* \mid T_j\right)}\right) \tag{24}$$

似然函数：

$$p\left(\boldsymbol{y} \mid T_j, M_j, \sigma^2\right) \propto \prod_{i=1}^{n} \exp\left(-\frac{w_i\left(y_i - f\left(\boldsymbol{x}_i\right)\right)^2}{2\sigma^2}\right) \tag{25}$$

$$p\left(\boldsymbol{y} \mid T_j^*, M_j^*, \sigma^2\right) \propto \prod_{i=1}^{n} \exp\left(-\frac{w_i\left(y_i - f^*\left(\boldsymbol{x}_i\right)\right)^2}{2\sigma^2}\right) \tag{26}$$

其中，f^* 为基于新树 T_j^* 的预测值；先验项 $p\left(T_j\right)$ 为树结构的先验概率；提议分布 $q\left(T_j^* \mid T_j\right)$ 为从旧树 T_j 生成新树 T_j^* 的概率。

(2) 叶节点 M_j 参数的条件后验分布

假设叶节点 μ_{jk} 的先验为正态分布，那么在给定树结构 T_j 和权重 w_i 的条件下，叶节点参数 μ_{jk} 的后验分布仍为正态分布，其中权重控制不同观测值对叶节点更新的影响强度。由此，叶节点参数的后验均值不再对应条件均值，而是自然收敛于目标 Expectile。考虑第 j 棵树的叶节点 k，记 S_{jk} 为落入第 j 棵树第 k 个叶节点的观测索引集合，其参数为 μ_{jk}，残差为 $r_i = y_i - \sum_{k \neq j} g_k\left(x_i\right)$。具体推导如下：

先验：

$$\mu_{jk} \sim N\left(0, \sigma_\mu^2\right) \tag{27}$$

似然函数：

$$p\left(\{r_i\} \mid \mu_{jk}, \sigma^2\right) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i \in S_{jk}} w_i\left(r_i - \mu_{jk}\right)^2\right) \tag{28}$$

联合分布：

$$\begin{aligned} p\left(\mu_{jk} \mid \cdot\right) &\propto p\left(\mu_{jk}\right) p\left(\{r_i\} \mid \mu_{jk}, \sigma^2\right) \\ &\propto \exp\left(-\frac{\mu_{jk}^2}{2\sigma_\mu^2}\right) \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{i \in S_{jk}} w_i\left(r_i - \mu_{jk}\right)^2\right) \\ &\propto \exp\left(-\frac{\mu_{jk}^2}{2}\left(\frac{1}{\sigma_\mu^2} + \frac{\sum w_i}{\sigma^2}\right) + \mu_l \cdot \frac{\sum w_i r_i}{\sigma^2}\right) \end{aligned} \tag{29}$$

后验分布：

$$\mu_{jk} \mid \cdots \sim N\left(\frac{\sum_{i \in S_{jk}} w_i r_i}{\sum_{i \in S_{jk}} w_i + \frac{\sigma^2}{\sigma_\mu^2}}, \frac{\sigma^2}{\sum_{i \in S_{jk}} w_i + \frac{\sigma^2}{\sigma_\mu^2}}\right) \tag{30}$$

(3) 更新方差 σ^2

假设方差先验为逆伽马分布，那么在给定当前树结构与权重的条件下，误差项方差的条件后验分布仍为逆伽马分布，其尺度参数由加权残差平方和更新。该更新方式反映了 Expectile 损失下的残差离散程度，并保持了计算上的共轭性与稳定性。具体推导如下：

先验：$\sigma^2 \sim \text{Inverse-Gamma}(a,b)$，即其概率密度函数为：

$$p(\sigma^2)=\frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right) \tag{31}$$

似然函数可写为：

$$(Y\mid F,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\left(\frac{\sigma^2}{w_i}\right)}}\exp\left(-\frac{w_i(y_i-f(x_i))^2}{2\sigma^2}\right) \tag{32}$$

联合分布：

$$\begin{aligned} p(\sigma^2\mid Y) &\propto p(Y\mid F,\sigma^2)p(\sigma^2) \\ &\propto (\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right)\cdot(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{\sum_{i=1}^{n}w_i\epsilon_i^2}{2\sigma^2}\right) \\ &\propto (\sigma^2)^{-\left(a+\frac{n}{2}+1\right)}\exp\left(-\frac{b+\frac{\sum_{i=1}^{n}w_i\epsilon_i^2}{2}}{\sigma^2}\right) \end{aligned} \tag{33}$$

后验分布为逆伽马分布：

$$\sigma^2\mid\cdots\sim\text{Inverse-Gamma}\left(a+\frac{n}{2},b+\frac{1}{2}\sum_{i=1}^{n}w_i\epsilon_i^2\right) \tag{34}$$

4. 实证分析

4.1. 数据来源

本文使用法国第三方责任保险数据集 freMTPL2freq 与 freMTPL2sev 进行实证分析。两数据集通过保单 ID 进行匹配，并计算每张保单在保险期间内的累计索赔金额，最终构建包含 677,991 个观测值和 12 个变量的综合数据集。其中，发生索赔的保单共 24,944 条。主要解释变量包括车辆与驾驶员特征(如车辆品牌、燃料类型、车辆与驾驶员年龄、奖励惩罚系数等)以及区域特征变量，响应变量分别为是否出险、累计索赔额及风险附加指标。主要解释变量包括车辆与驾驶员特征(如车辆品牌、燃料类型、车辆与驾驶员年龄、奖励惩罚系数等)以及区域特征变量，响应变量分别为是否出险、累计索赔额及风险暴露指标。

4.2. 第一阶段：出险概率模型

针对是否发生索赔样本分布不均衡的问题，采用 ROSE 包中的过采样与欠采样相结合方法对原始数据进行平衡处理，生成规模为 60,000 的平衡样本集。选取车辆、驾驶员及区域特征作为自变量，是否发生索赔作为因变量，将样本按 7:3 划分为训练集与测试集。

基于统一的数据划分，通过 R 软件分别构建 Logit 回归、随机森林、XGBoost 以及 BART 模型进行对比分析。各机器学习模型均通过交叉验证与网格搜索进行参数调优，并以 F1 分数为主要评价指标，兼

顾 AUC、准确率、精确率与召回率等综合评估指标，选取最优参数组合，并在测试集上，比较不同模型在保险索赔概率建模中的预测性能与稳定性表现，结果如表 1 所示。测试集结果表明，XGBoost 模型在 AUC、AUPRC 及 F1 等指标上均取得最优表现，显示出较强的区分能力与稳定性，因此在后续定价流程中作为出险概率预测模型。

Table 1. Performance comparison of claim occurrence probability models in auto insurance
表 1. 基于车险出险概率模型各指标比较

Model	AUC	AUPRC	准确率	精确率	召回率	F1 分数
Logit	0.665	0.632	0.601	0.565	0.890	0.691
BART	0.698	0.663	0.617	0.574	0.911	0.705
RandomForest	0.788	0.777	0.689	0.633	0.904	0.745
XGBoost	**0.793**	**0.774**	**0.697**	**0.641**	**0.902**	**0.749**

4.3. 第二阶段：累计索赔额模型

在发生索赔的条件下，对累计索赔额大于零的样本进行建模分析，其样本分布图如图 2 所示。考虑到索赔金额具有明显的右偏和厚尾特征，对其进行对数变换后作为响应变量。样本同样按 7:3 划分为训练集与测试集。

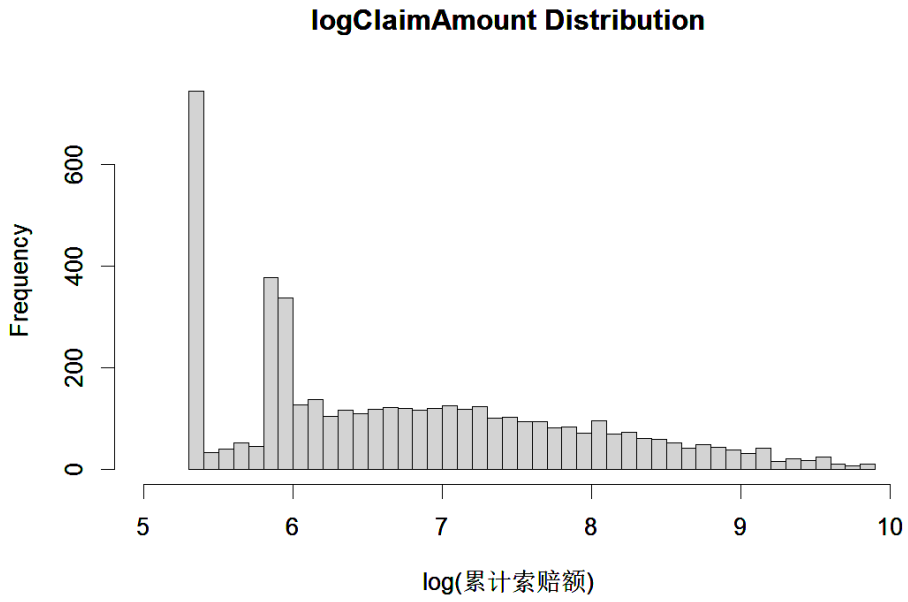

Figure 2. Distribution of log-transformed cumulative claim amounts
图 2. 对数累计索赔额分布

通过 R 软件分别构建随机森林、XGBoost 以及 BART 模型进行对比分析。模型训练前，对数据进行统一的数值化与预处理，以保证各模型输入格式的一致性。在参数选择阶段，基于训练集采用交叉验证方法对关键超参数进行调优，通过在预设参数空间内搜索不同组合，综合比较模型在验证集上的预测误差，从而确定最优参数配置。模型训练完成后，在独立测试集上进行预测，分别计算 MSE、RMSE 和 MAE 等回归评价指标，以系统评估不同模型的预测性能，结果如表 2 所示。从模型表现来看，BART 模型在各项指标上均优于其他模型，表明其在刻画索赔金额的非线性结构和预测精度方面具有明显优势。

Table 2. Performance comparison of models for cumulative claim amount prediction
表 2. 基于车险累计索赔额模型各指标比较

Model	MSE	RMSE	MAE
RandomForest	1.351163	1.162395	0.7958445
XGBoost	1.730985	1.315669	0.9437754
BART	**1.273390**	**1.128446**	**0.7668005**

4.4. 风险附加模型

在前两阶段模型基础上，进一步构建 BEART 模型，用于刻画索赔金额的尾部风险特征。通过固定分位数 $\tau = 0.8$，通过 MCMC 迭代在每一轮更新中根据残差符号动态调整权重，并对误差方差进行贝叶斯更新，从而实现非对称损失下的加权拟合。基于后验样本对训练集和测试集进行预测，并采用 WMSE 与 WMAD 进行评估，结果如表 3 所示。在各模型中，BEART 模型在两项指标上均表现最优，显著优于其他方法，体现出其在捕捉车险索赔数据右偏厚尾特征和刻画尾部风险方面的优势，更适用于高风险个体的精细化定价与风险管理。

Table 3. Performance comparison of risk loading models
表 3. 风险附加模型各指标比较

Model	WMAD	WMSE
RandomForest	0.3921060	0.5940003
XGBoost	0.4604460	0.7806970
BART	0.3708921	0.5424180
BEART	**0.2604471**	**0.4444054**

5. 结论

本文围绕车险风险保费定价中的关键建模问题，构建并验证了一套分阶段的建模框架，重点关注风险识别能力与尾部风险刻画效果。在出险概率、累计索赔额及风险附加三个核心环节中，分别引入并比较多种统计与机器学习模型，系统评估其在不均衡数据、非线性关系及厚尾分布情形下的预测表现。实证结果表明，基于贝叶斯期望分位数加性回归树的风险附加模型在尾部风险识别方面具有明显优势，能够更有效地刻画高风险保单的损失特征。

需要指出的是，风险保费的最终厘定通常还需结合保险公司的经营目标、监管约束以及风险偏好等外生因素，其结果并非单一模型输出所能直接确定。鉴于此，本文未对各阶段模型进行数值层面的直接合成，而是侧重于验证各子模型在预测精度与风险刻画能力方面的有效性，从方法论层面为后续完整风险保费定价方案的构建提供可行的建模模块与技术路径。未来研究可在此基础上，进一步引入经营约束与定价规则，对风险附加系数及分位数水平进行情景化设定，从而实现模型结果与实际定价决策的有机结合。

参考文献

[1] 杨亮, 孟生旺. 基于分位回归的风险保费预测[J]. 统计与信息论坛, 2016, 31(9): 83-88.

[2] Heras, A., Moreno, I. and Vilar-Zanón, J.L. (2018) An Application of Two-Stage Quantile Regression to Insurance Ratemaking. *Scandinavian Actuarial Journal*, **2018**, 753-769. https://doi.org/10.1080/03461238.2018.1452786

[3] Yang, L., Li, Z. and Meng, S. (2020) Risk Loadings in Classification Ratemaking. arXiv: 2002.01798.

[4] 蔡超, 黄聪聪, 董皓天. 分位数回归提升树模型及应用[J]. 系统科学与数学, 2022, 42(5): 1216-1233.

[5] Cai, C., Dong, H. and Wang, X. (2022) Expectile Regression Forest: A New Nonparametric Expectile Regression Model. *Expert Systems*, **40**, e13087. https://doi.org/10.1111/exsy.13087

[6] Jiang, C., Jiang, M., Xu, Q. and Huang, X. (2017) Expectile Regression Neural Network Model with Applications. *Neurocomputing*, **247**, 73-86. https://doi.org/10.1016/j.neucom.2017.03.040

[7] Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, **4**, 266-298. https://doi.org/10.1214/09-aoas285

[8] 郭哲琦, 高苏浩. 基于两阶段 Expectile 回归的风险保费定价[J]. 统计与决策, 2024, 40(3): 162-167.

[9] Newey, W.K. and Powell, J.L. (1987) Asymmetric Least Squares Estimation and Testing. *Econometrica*, **55**, 819-847. https://doi.org/10.2307/1911031

[10] 曹桃云, 张日权. 非对称误差分布的贝叶斯累加回归树模型研究及应用[J]. 系统科学与数学, 2022, 42(11): 3119-3133.

[11] 金娇, 管思晴, 扈灵嫣, 等. Expectile 回归模型的贝叶斯统计推断研究[J]. 统计学与应用, 2021, 10(5): 929-939.