

基于随机森林和Cox模型的肺癌生存时间分析与预测

赵司健

云南财经大学统计与数学学院, 云南 昆明

收稿日期: 2024年7月8日; 录用日期: 2024年7月29日; 发布日期: 2024年8月12日

摘要

本研究旨在通过随机生存森林和Cox比例风险模型, 分析并预测肺癌患者的生存时间。研究数据来源于R语言中的survival包中的cancer数据集。首先, 采用随机森林方法进行变量选择, 结果显示, 性别和体能状态是对生存时间具有显著影响的关键变量。接着, 本文使用Cox比例风险模型进一步分析了上述变量对生存时间的影响。结果显示, 体能状态评分越高, 死亡风险越大, 而女性患者的生存时间相对较长但统计显著性较低。Cox比例风险模型的分析表明, 模型在区分生存时间上的能力较好, 且模型整体显著。为了直观展示不同风险组的生存概率差异, 绘制了生存曲线, 结果表明, 高风险组的生存概率显著低于低风险组。通过绘制ROC曲线并计算AUC值, 发现模型在区分高低风险患者方面具有中等的预测能力。此外, Bootstrap方法验证了模型的稳定性, 性别和体能状态的系数在多次抽样中的估计值较为一致。模型贡献解释中通过Shapley值进一步验证了性别和体能状态是预测生存时间的重要指标, 确认了它们在模型中的关键作用。综上所述, 本研究通过系统的变量选择、模型分析和多种评估方法, 揭示了性别和体能状态对肺癌患者生存时间的显著影响, 并验证了模型的稳健性和有效性, 为临床实践中预测患者预后提供了重要的参考依据。

关键词

肺癌, 生存时间预测, 随机森林, Cox比例风险模型, ROC曲线, Bootstrap方法

Analysis and Prediction of Lung Cancer Survival Time Based on Random Forest and Cox Model

Sijian Zhao

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan

Abstract

The aim of this study was to analyze and predict the survival time of lung cancer patients by means of random survival forests and Cox proportional risk models. The study data were obtained from the cancer dataset in the survival package in R language. First, the random forest method was used for variable selection, and the results showed that gender and physical status were the key variables that had a significant effect on survival time. Then, this paper further analyzes the effects of the above variables on survival time using Cox proportional risk model. The results showed that the higher the physical status score, the higher the risk of death, while female patients had relatively longer but less statistically significant survival times. The analysis of the Cox proportional risk model showed the model's ability to discriminate between survival times was better and the model was overall significant. Survival curves were plotted to visualize the difference in survival probability between different risk groups, and the results showed that the survival probability of the high-risk group was significantly lower than that of the low-risk group. By plotting the ROC curve and calculating the AUC value, the model was found to have moderate predictive ability in distinguishing between high- and low-risk patients. In addition, the Bootstrap method verified the stability of the model, and the coefficients for gender and physical status were more consistent in their estimates across multiple samples. The model contribution interpretation was further validated by the Shapley value that gender and physical fitness status are important predictors of survival time, confirming their key role. In summary, this study revealed the significant effects of gender and physical status on the survival time of lung cancer patients through systematic variable selection, model analysis and multiple assessment methods, and verified the robustness and validity of the model, which provides an important reference for predicting patients' prognosis in clinical practice.

Keywords

Lung Cancer, Survival Time Prediction, Random Forest, Cox Proportional Risk Model, ROC Curve, Bootstrap Method

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

肺癌是全球范围内最常见和致命的癌症之一，其高发病率和死亡率引起了广泛关注。据世界卫生组织统计，2020 年全球有超过 220 万例新发肺癌病例，且肺癌是导致癌症相关死亡的主要原因，约占所有癌症死亡的 18% [1]。这不仅给患者及其家庭带来了巨大的痛苦和经济负担，也对公共卫生系统提出了严峻挑战。根据国际癌症研究机构的数据，2022 年肺癌仍然是最常见的癌症类型，并且是癌症死亡的首要原因，预计造成 180 万人死于肺癌[2]。对于肺癌患者的生存情况成为了研究的主流方向，对于处理生存数据，Cox 比例风险模型是一个强有力的工具。Frank Emmert-Streib 等学者详细讨论了 Cox 比例风险模型以及检验比例风险(PH)假设的方法，对于 PH 假设不成立的情况，利用分层 Cox 模型。并且通过使用统计编程语言 R 对肺癌患者数据进行了分析和研究[3]。目前，利用机器学习算法进行肺癌预测和诊断也成为一

个重要的研究方向。机器学习技术，特别是分类算法，因其强大的数据处理和模式识别能力，成为研究者们关注的焦点。朴素贝叶斯、支持向量机(SVM)、决策树和随机森林等算法被广泛应用于医学诊断领域，通过分析患者的症状、生活习惯以及其他相关数据，这些算法能够有效预测肺癌的发生[4] [5]。具体来说，朴素贝叶斯算法因其简单高效，适用于初步筛查；支持向量机以其在高维空间中处理非线性分类问题的能力，成为精确诊断的有力工具；决策树算法则通过树状结构清晰地展示决策过程，便于解释和应用；随机森林结合了多个决策树的优点，提高了预测的准确性和鲁棒性[3]。这些算法的性能比较和优化，是当前研究的一个重要方向。

此外，数据集的选择和处理也在肺癌预测中扮演着重要角色。例如，UCI 机器学习库中的癌症数据集被广泛用于模型评估和优化。这些数据集包含了大量真实世界的患者信息，为模型的训练和测试提供了丰富的数据支持[6]。通过对这些数据集的深入分析，研究人员能够提取出关键特征，提高模型的预测性能。

2. 数据描述性分析

本文使用的数据集为 R 语言中 survival 包中的 cancer 数据集，下面对该数据的所有变量进行分析。

1) 变量说明

本文所用到的数据集变量名称及说明在表 1 中进行了详细展示。

Table 1. Variable description

表 1. 变量说明

变量名称	说明
inst	研究机构编号
time	生存时间(天)
status	生存状态(1 = 删失, 2 = 死亡)
age	年龄(岁)
sex	性别(1 = 男性, 2 = 女性)
ph.ecog	ECOG 绩效状态(0 = 完全正常, 5 = 患者死亡)
ph.karno	Karnofsky 评分(医师评估的患者功能状态)
pat.karno	Karnofsky 评分(患者评估的功能状态)
meal.cal	每日摄入卡路里(卡路里)
wt.loss	体重减轻情况(磅)

从表 1 中可以了解到关于变量的基本性质及其含义，为后续的研究提供了有用信息。

2) 变量相关性可视化

图 1 显示了多个变量之间的关系：时间、性别、年龄、ECOG 体能状态评分、卡氏体能状态评分、患者卡氏体能状态评分、每日卡路里摄入量和体重减轻。ECOG 体能状态评分和卡氏体能状态评分之间存在强负相关(-0.8)，表明当一个评分增加时，另一个评分显著下降。卡氏功能状态评分和患者卡氏体能状态评分之间有中等正相关(0.51)，体重减轻和性别之间有轻微负相关(-0.13)。整体来看，ECOG 体能状态评分与生存时间呈较强负相关，卡氏体能状态评分与生存时间呈轻微正相关，性别和患者的卡氏体能状态评分对生存时间也有影响。这些发现为未来的研究和模型构建提供了重要依据，提示在预测肺癌患者生存时间和制定治疗方案时需重点考虑这些变量的影响，并探讨改善患者营养状况和身体机能状态的方法。

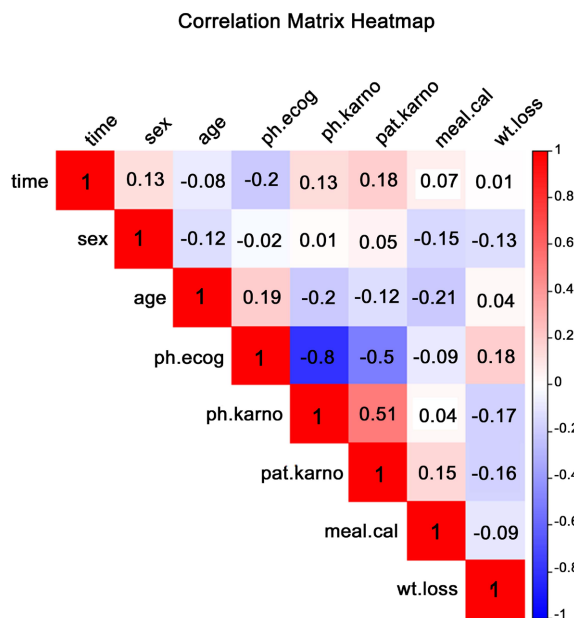


Figure 1. Variable correlation heat map

图 1. 变量相关性热力图

3. 模型建立

3.1. 数据预处理方法简述

在本研究中，本文对多变量生存数据进行了详尽的预处理。首先，导入原始数据，删除非核心列，仅保留与生存分析相关的变量。随后，将状态变量重新编码为二元格式。最后，对数据进行了标准化处理，通过中心化和缩放调整变量尺度，提高模型稳定性、解释能力和预测精度。处理好的数据集随机分为训练集和测试集，确保模型能有效预测新数据，避免过拟合。这些步骤确保了数据的准确性和可靠性，为高效生存分析模型的构建奠定了坚实基础。

3.2. 模型构建

随机森林是一种集成学习方法，通过构建多个决策树并将其结果进行组合来提高模型的性能和稳定性。对于生存分析，随机森林可以扩展为生存随机森林。

1) 随机森林模型

随机森林模型是一种基于决策树的集成学习方法，通过构建多个决策树来提高模型的稳定性和预测准确性，数学表达式为[7]：

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$$

其中， $\hat{y}$ 是预测值， N 为决策树的数量， $T_i(x)$ 为第 i 棵决策树对输入变量 x 的预测值。

2) 生存随机森林模型

对于生存分析，随机森林模型可以表示为[8]：

$$\hat{h}(t|X) = \frac{1}{B} \sum_{b=1}^B \hat{h}_b(t|X)$$

其中， $\hat{h}(t|X)$ 是个体在时间 t 的预测风险函数，给定协变量 X ； B 是森林中树的数量； $\hat{h}_b(t|X)$ 是第 b 棵

树的预测风险函数。

3) 树的构建

在构建每棵生存随机森林决策树时，首先从原始训练数据集中有放回地随机抽取一个子集，然后在每个节点分裂时随机选择一个特征子集，选择使得生存差异最大的特征及其分裂点，通常通过最大化对数秩检验统计量来确定。分裂过程持续到节点中的样本数少于预设的最小值或满足其他预设的终止条件时停止。

节点分裂的数学推导，设定在节点 m 处的样本为 $\{(t_i, \delta_i, X_i)\}_{i \in N_m}$ ，其中 N_m 表示节点 m 的样本索引集合。为了选择最佳分裂点 s 和特征 j ，最大化对数秩检验统计量[8] [9]：

$$Z_{m,j}(s) = \frac{\sum_{i \in N_m} (\delta_i - E_i) I(X_{ij} \leq s)}{\sqrt{\sum_{i \in N_m} V_i I(X_{ij} \leq s)}}$$

其中， E_i 是在 $X_{ij} \leq s$ 时，事件的期望数。 V_i 是在 $X_{ij} \leq s$ 时，事件的方差。通过遍历所有的特征和分裂点，选择使得 $|Z_{m,j}(s)|$ 最大的 s 和 j 。

4) 超参数调优

生存随机森林模型的超参数包括每个节点分裂时随机选择的特征数量、终止节点分裂的最小样本数以及森林中树的数量。通过网格搜索的方法，对这些超参数进行调优，选择使得模型性能最佳的参数组合。

5) LASSO 回归

在随机森林模型的基础上，本文应用了 LASSO 回归进行变量选择。LASSO 回归通过引入 L1 正则化项，实现自动变量选择和模型简化，保留对生存时间影响显著的变量，减少模型的复杂性，防止过拟合，从而提高模型的解释性和预测性能。

LASSO 数学公式如下[7]：

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

其中， y 是第 i 个样本的响应变量， x_{ij} 是第 i 个样本的第 j 个预测变量， β_0 是截距项， β_j 是回归系数， λ 是正则化参数，用于控制模型的稀疏性。

6) 预测与评估

a) 预测

对于新的数据点 X ，生存随机森林的预测是所有树的预测结果的平均[8]：

$$\hat{h}(t|X) = \frac{1}{B} \sum_{b=1}^B \hat{h}_b(t|X)$$

b) 评估

使用 ROC 曲线和 AUC 值来评估模型在特定时间点的预测性能：

假设我们在时间点 t 处计算 ROC 曲线，则 AUC 的定义为[7]：

$$\text{AUC} = \int_0^1 \text{TPR}(u) d(\text{FPR}(u))$$

其中：假阳性率(FPR)和真阳性率(TPR)分别定义为[7]：

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

通过计算不同阈值下的 FPR 和 TPR，并绘制 ROC 曲线，计算 AUC 值。

7) 生存分析模型的稳定性评估和解释

a) 模型稳定性评估

为了评估生存分析模型的稳定性，本文使用 Bootstrap 方法，这是一种从数据集中有放回地进行多次抽样的技术。具体步骤包括：从原始数据集中随机抽取样本，形成一个与原始数据集大小相同的子集(有放回)，对每个抽样子集拟合 Cox 比例风险模型并记录模型参数。通过重复上述过程多次，可以得到多个模型参数估计，并通过计算这些参数估计的均值、标准差等统计量来评估模型参数的稳定性。

数学推导如下：

设原始数据集为 D ，大小为 n 。每次 Bootstrap 抽样生成的数据集为 D^* ，对每个数据集拟合模型，得到参数估计 $\hat{\beta}^*$ 。

参数稳定性的估计过程如下：

① 抽样与拟合模型[10]：

$$D_b^* = \left\{ (t_i^*, \delta_i^*, X_i^*) \right\}_{i=1}^n$$

$$\hat{\beta}_b = \text{coef} \left(\text{coxph} \left(\text{Surv}(t_i^*, \delta_i^*) \sim X_i^*, \text{data} = D_b^* \right) \right)$$

其中， t_i^* 表示第 i 个样本的生存时间， δ_i^* 是第 i 个样本的删失指示符， $t_i^* \text{Surv}(t_i^*, \delta_i^*)$ 表示生存时间和删失指示符作为因变量， X_i^* 作为自变量， coxph 是 Cox 比例风险模型的函数， coef 提取模型的系数(即估计的参数)。

② 重复抽样 R 次：

$$\hat{\beta}^{(1)}, \dots, \hat{\beta}^{(R)}$$

③ 计算参数估计的均值和标准差：

$$\bar{\beta} = \frac{1}{R} \sum_{b=1}^R \hat{\beta}_b$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{R-1} \sum_{b=1}^R (\hat{\beta}_b - \bar{\beta})^2}$$

b) 模型解释

SHAP (Shapley Additive Explanations) 值是一种基于博弈论的模型解释方法，用于量化每个特征对模型预测的贡献。以下是 SHAP 值计算的数学推导[11]：

Shapley 值最初来自合作博弈论，用于分配合作收益，对每个特征 i ，其 Shapley 值定义为：

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

其中， N 是所有特征的集合， S 是 N 的一个子集，不包含特征 i ， $v(S)$ 是特征集合 S 对应的预测值。

SHAP 值在模型解释中的应用：对于每个样本 x ，SHAP 值表示特征 i 对模型输出的贡献。模型输出可以表示为所有特征 SHAP 值的加和：

$$f(X) = \phi_0 + \sum_{i=1}^M \phi_i$$

其中， ϕ_0 是基准值(通常是训练数据的平均预测值)。

计算每个特征对所有可能特征组合的贡献，取所有组合的加权平均。使用数值方法，如蒙特卡罗模拟或近似算法，加速计算过程。通过 Bootstrap 方法进行模型稳定性评估可以帮助我们了解模型系数的波动范围和置信区间，而使用 SHAP 值进行模型解释可以直观地量化每个特征对模型预测结果的影响。这两种方法结合，可以有效提高模型的解释性和可信度。

4. 实证分析

4.1. 变量选择与模型参数设置

在本研究中，首先使用随机生存森林方法来进行变量选择，并对模型的预测性能进行了详细评估。变量选择的目标是从多个潜在预测变量中识别出对生存时间具有显著影响的关键变量。在本研究中，使用了 R 语言中 survival 包中的 cancer 数据集进行实证分析。该数据集广泛应用于生存分析的研究和教学中，包含了肺癌患者的生存数据及相关临床变量。

通过运行 R 语言得到的结果如表 2 所示。

Table 2. Variable selection results
表 2. 变量选择结果

输出参数名称	对应的值(类型)
family	surv
var. selection	Variable Hunting (VIMP)
conservativeness	medium
dimension	7
sample size	104
K-fold	5
no. reps	50
nstep	1
ntree	500
nsplit	10
mvars	2
nodesize	2
refitted forest	TRUE
model size	1.68 ± 0.4712
PE (K-fold)	45.9126 ± 8.3596
Top variables	rel.freq
sex	30
ph.ecog	26

由表 2 可知，“Variable Hunting (VIMP)”方法进行变量选择，结果显示性别(sex)和体能状态(ph.ecog)在多次随机生存森林的训练中分别出现 30 次和 26 次，表明这两个变量对生存时间具有重要的预测作用。模型参数设置如下：变量选择保守度为中等，数据维度为 7，样本量为 104，K 折交叉验证为 5，重复次

数为 50，步数为 1，决策树数量为 500，分裂次数为 10，变量数为 2，节点大小为 2，模型重新拟合，模型规模为 1.68 ± 0.4712 ，预测误差(K 折)为 45.9126 ± 8.3596 。这些参数的选择和设置确保了模型在训练过程中具有良好的稳定性和泛化能力，从而能够在新数据上保持较高的预测性能。

4.2. Cox 比例风险模型总结

为了进一步探讨变量对生存时间的影响，本文采用了 Cox 比例风险模型进行分析。Cox 模型是一种常用的生存分析方法，其优势在于能够处理多变量同时对生存时间的影响。

1) Cox 比例风险模型的回归结果如表 3 所示。

Table 3. Regression results

表 3. 回归结果

变量	coef	exp (coef)	se (coef)	z	Pr (> z)	exp (coef) lower 0.95	exp (coef) upper 0.95
sex	-0.4004	0.6701	0.2170	-1.845	0.0651	0.4379	1.025
ph.ecog	0.5759	1.7788	0.1405	4.098	4.17E-05	1.3505	2.343

由表 3 可知，性别(sex)的系数为-0.4004，HR = 0.6701， $p = 0.0651$ ，体能状态(ph.ecog)的系数为 0.5759，HR = 1.7788， $p = 4.17e-05$ ，从系数的符号和显著性水平可以看出，体能状态对生存时间有显著的正向影响，即体能状态评分越高，死亡风险越大；性别对生存时间也有一定影响，女性的生存时间相对较长，但该结果在统计上仅接近显著($p = 0.0651$)。

Table 4. Model statistical results

表 4. 模型统计结果

统计量	值
Concordance	0.654 (se = 0.032)
Likelihood ratio test	19.48 on 2 df, $p = 6e-05$
Wald test	19.61 on 2 df, $p = 6e-05$
Score (logrank) test	20.07 on 2 df, $p = 4e-05$

由表 4 可知，Cox 模型的 Concordance 指数为 0.654，似然比检验、Wald 检验及评分检验均显示模型整体显著， p 值均小于 0.001，这些结果表明 Cox 模型在区分生存时间上的能力较好，且模型整体显著。

4.3. 生存分析图

绘制生存曲线，直观地展示高风险和低风险组的生存概率差异。

结果解读：图 2 展示了低风险组和高风险组患者的生存概率随时间的变化。可以看出，高风险组的生存概率显著低于低风险组。从图中可以观察到，随着时间的推移，高风险组的生存概率迅速下降，而低风险组的生存概率则相对较高且下降速度较慢。这一结果与 Cox 模型分析的结论一致，进一步验证了体能状态和性别对生存时间的影响。

4.4. 计算 ROC 曲线和 AUC 值

为了评估模型在区分高风险和低风险患者方面的能力，本文绘制了 ROC 曲线并计算了 AUC 值。

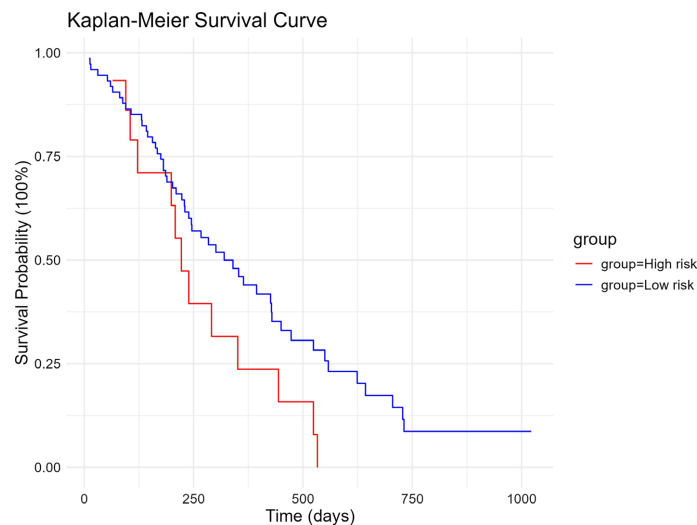


Figure 2. Survival curve
图 2. 生存曲线图

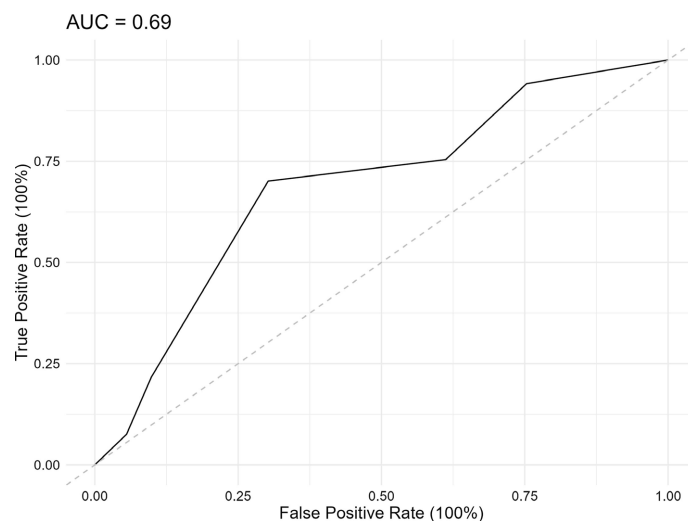


Figure 3. ROC curve
图 3. ROC 曲线图

图 3 展示了模型的分类性能，横轴表示假阳性率(False Positive Rate)，纵轴表示真阳性率(True Positive Rate)。ROC 曲线的面积(AUC)为 0.69，这表明模型在区分高风险和低风险患者方面具有中等的预测能力，模型在一定程度上能够区分高风险和低风险患者。总体而言，图 3 表明模型在区分高低风险病人任务中的表现良好。

4.5. 模型稳定性评估

模型稳定性评估采用 Bootstrap 方法，其分析结果如表 5 所示。

由表 5 可知，性别(sex)的系数原始值为-0.4004，偏差为-0.0239，标准误为 0.2259；体能状态(ph.ecog)的系数原始值为 0.5759，偏差为 0.0298，标准误为 0.1708，这些结果进一步验证了模型的稳定性和可靠性，表明通过不同数据抽样方法得到的系数估计值较为一致，性别和体能状态变量在多次抽样中的系数估计值接近原始数据的估计值，说明模型结果具有较高的稳健性。

Table 5. Bootstrap analysis results
表 5. Bootstrap 分析结果

统计量	原始值	偏差	标准误
t_1^*	-0.4003683	-0.02390182	0.2258895
t_2^*	0.5759396	0.02977760	0.1708331

4.6. 模型贡献解释

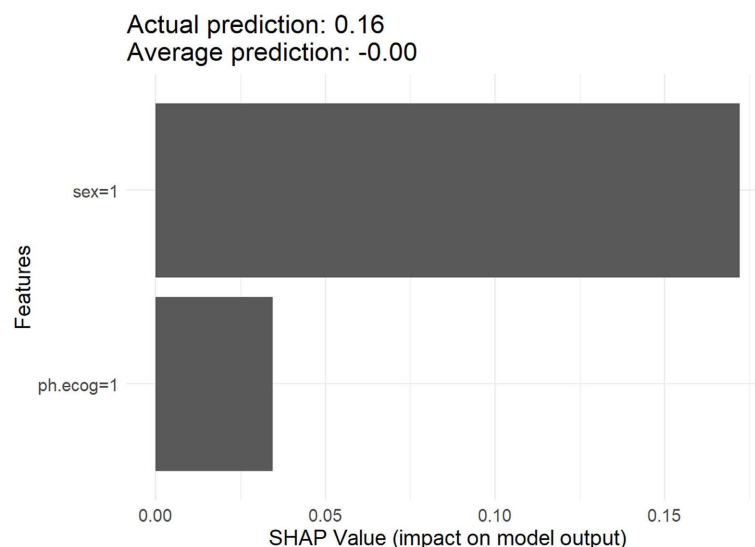


Figure 4. Variable importance analysis
图 4. 变量重要性分析

图 4 展示了变量的重要性排序，实际预测值为 0.16，而平均预测值接近于 0，这表明特征的组合(性别和体能状态)对预测结果有显著影响。具体来说，这两个特征显著提升了模型的预测结果，远高于基线预测值。通过 Shapley 值的分解，我们可以看到性别和体能状态分别贡献了约为 0.17 和 0.03 的 Shapley 值，进一步验证了它们在变量选择过程中的重要性。这一结果强调了在实际应用中，性别和体能状态是预测患者生存时间的关键指标，它们对模型决策起到了至关重要的作用，从而使得预测结果更加精准和可靠。

5. 结论和建议

5.1. 结论

本研究结合随机生存森林和 Cox 比例风险模型，识别并分析了肺癌患者生存时间的关键影响因素。通过随机生存森林重要变量选择方法，发现性别(sex)和体能状态(ph.ecog)是对生存时间影响最显著的变量。体能状态评分越高，死亡风险越大，性别对生存时间也有一定影响，女性的生存时间相对较长。Cox 模型的 Concordance 指数为 0.654，表明模型在区分生存时间上的较好能力，且似然比检验、Wald 检验及评分检验均显示模型整体显著，p 值均小于 0.001。通过非参数 bootstrap 方法对 Cox 模型的系数进行估计，验证了模型的稳定性和可靠性，系数估计值在多次抽样中的一致性较高。在模型贡献解释中，Shapley 值分别约为 0.17 和 0.03，显示出性别和体能状态对模型预测结果的重要贡献，进一步确认它们是影响生存时间的主要因素。这些发现为临床决策和个性化治疗提供了重要参考。未来研究应进一步优化模型，提

高预测精度，并探索更多潜在的影响因素，以期更好地服务于患者管理和治疗策略的制定。

5.2. 建议

基于本研究结果，提出以下三点建议，以进一步提升模型在生存分析中的应用价值：

1) 数据质量和完整性

建议在未来的研究中扩大样本量，涵盖不同人群和疾病阶段，以增强模型的鲁棒性和广泛适用性。采用更先进的缺失值填补方法，如多重插补，以提高数据的完整性和模型的精度。

2) 模型优化与验证

进一步优化模型的超参数，如决策树的数量、每次分裂时选择的变量数和节点的最小样本数。可以使用网格搜索或贝叶斯优化等方法，找到最佳的参数组合。采用更严格的交叉验证方法，如 k 折交叉验证，进一步验证模型的稳健性和泛化能力，确保模型在不同数据划分下的稳定性。与其他模型(如加权生存模型、支持向量机和神经网络等)进行比较，找出在不同应用场景中表现最优的模型。

3) 实际应用

将模型集成到临床决策支持系统中，实时预测患者的生存概率，辅助医生制定个体化的治疗方案，提高临床决策的准确性和效率。利用模型预测结果，对高风险患者进行重点监控和管理，提供个性化的干预措施，延长患者的生存时间，提高生活质量。

参考文献

- [1] World Health Organization (2022) Cancer. <https://www.who.int/news-room/fact-sheets/detail/cancer>
- [2] International Agency for Research on Cancer (IARC) (2024) New Report on Global Cancer Burden in 2022 by World Region and Human Development Level. <https://www.iarc.who.int/news-events/new-report-on-global-cancer-burden-in-2022-by-world-region-and-human-development-level/>
- [3] Emmert-Streib, F. and Dehmer, M. (2019) Introduction to Survival Analysis in Practice. *Machine Learning and Knowledge Extraction*, **1**, 1013-1038. <https://doi.org/10.3390/make1030058>
- [4] Radhika, P.R., Nair, R.A.S. and Veena, G. (2019). A Comparative Study of Lung Cancer Detection Using Machine Learning Algorithms. 2019 *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, 20-22 February 2019, 1-4. <https://doi.org/10.1109/icecct.2019.8869001>
- [5] Rajini, A. and Jabbar, M.A. (2021) Lung Cancer Prediction Using Random Forest. *Recent Advances in Computer Science and Communications*, **14**, 1650-1657. <https://doi.org/10.2174/2213275912666191026124214>
- [6] Anil Kumar, C., Harish, S., Ravi, P., SVN, M., Kumar, B.P.P., Mohanavel, V., *et al.* (2022) Lung Cancer Prediction from Text Datasets Using Machine Learning. *BioMed Research International*, **2022**, Article ID: 6254177. <https://doi.org/10.1155/2022/6254177>
- [7] Zhou, Z.-H. (2016) *Machine Learning*. Tsinghua University Press.
- [8] Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S. (2008) Random Survival Forests. *The Annals of Applied Statistics*, **2**, 841-860. <https://doi.org/10.1214/08-aoas169>
- [9] Klein, J.P. and Moeschberger, M.L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.
- [10] Xiao, Y.L. (2005) *Bootstrap-Based Inference for Cox's Proportional Hazards Analyses of Clustered Censored Survival Data*. Master's Thesis, McGill University.
- [11] Lundberg, S.M., Erion, G.G. and Lee, S.I. (2018) Consistent Individualized Feature Attribution for Tree Ensembles. arXiv: 1802.03888.