Operations Research and Fuzziology 运筹与模糊学, 2021, 11(4), 400-406
Published Online November 2021 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2021.114044

Hans汉斯

基于随机森林的保险费用影响因素分析

徐雪松

云南财经大学，云南 昆明

收稿日期：2021年9月23日；录用日期：2021年10月25日；发布日期：2021年11月1日

摘　要

随着人们的生活节奏越来越快，一部分人由于工作性质、工作强度的缘故，身心健康面临着重大挑战。吸烟、年龄上升，都会增加自身健康风险，而保险具有对冲这种风险的作用，因此购买保险成为人们抵消未来风险的一种手段。保险费用作为我们取得对未来的保障的费用，因此对保险费预测，不仅有助于我们认识不确定的未来，协助我们对未来的工作进行规划，而且有助于保险人对自己理赔风险的把控。本文运用随机森林模型预测保险费用，发现抽烟在很大程度上会增长保险费用，并且模型整体能解释保险费用变动的83.67%，拟合效果较好。

关键词

随机森林，变量重要性，R语言

Analysis of Influencing Factors of Insurance Expenses Based on Random Forest

Xuesong Xu

Yunnan University of Finance and Economics, Kunming Yunnan

Received: Sep. 23rd, 2021; accepted: Oct. 25th, 2021; published: Nov. 1st, 2021

Abstract

As the pace of people's lives becomes faster and faster, some people face major challenges on their physical and mental health due to the nature and intensity of their work. Smoking and increasing with age will increase their own health risks, and insurance has the effect of hedging this risk. Therefore, purchasing insurance has become a means for people to offset future risks. Insurance costs are the cost for us to obtain protection for the future, so forecasting insurance premiums will not only help us understand the uncertain future, plan our future work, but also help insurers

文章引用: 徐雪松. 基于随机森林的保险费用影响因素分析[J]. 运筹与模糊学, 2021, 11(4): 400-406.
DOI: 10.12677/orf.2021.114044

control their own claims risks. This paper uses the random forest model to predict insurance costs and finds that smoking will increase insurance costs to a large extent, and the model as a whole can explain 83.67% of the changes in insurance costs, and the fitting effect is good.

Keywords

Random Forest, Variable Importance, R Programming Language

1. 引言

随着时代的进步，保险在人们的日常生活中扮演着越来越重要的角色，人们购买保险的需求也越来越旺盛，但保险费用对消费者而言是否过多，是否超过自己能承受的范围，如何选择以缴纳合理的保险费，从哪些方面来评判自身需要缴纳多少保险费用，常常是消费者选择时的痛点，这也经常给消费者带来许多困扰[1]；同时，保险公司对消费者收取保险费也不是越多越好，收费越高，购买保险的人数越少；收费较低，会使自身收益受到损失。到底收取多少费用才能够提高自己的收益，对不同特征的消费者收取不同的保险费，通过消费者的哪些特征来判断，这些问题也给保险公司带来不小的困惑。因此针对保险费进行预测，利用数据分析探究影响保险费的主要因素，可以让消费者对缴纳保险费有一定规划，也为保险公司针对不同人群收取保险费提供了依据。

2. 对保险费样本数据的探索性分析

2.1. 保险费样本数据介绍

本文保险费样本数据总共包含 7 个变量，包括数值型和离散型变量。数值型变量有：age (年龄)、BMI (健康指数)、children (被保险家庭儿童数)、charges (保险费)；离散型变量有：sex (性别)、smoker (是否吸烟)、region (家庭方位)，详细情况见表 1。

Table 1. Independent variables names' explanation
表 1. 自变量名称解释

自变量名称	自变量含义解释	数值类型
Age	年龄	连续型数值
Sex	性别	二分类数值
BMI	健康指数	连续型数值
Children	被保险家庭儿童数	连续型数值
Smoker	是否吸烟	二分类数值
Region	家庭方位	四分类数值
Charges	保险费	连续型变量

2.2. 保险费样本数据探索性分析

从图 1 上看，charges (保险费)的偏度值为 1.51，是右偏分布，大部分数据分布在[0~17000]之间。说明大部分人的保险费集中在这个区间。

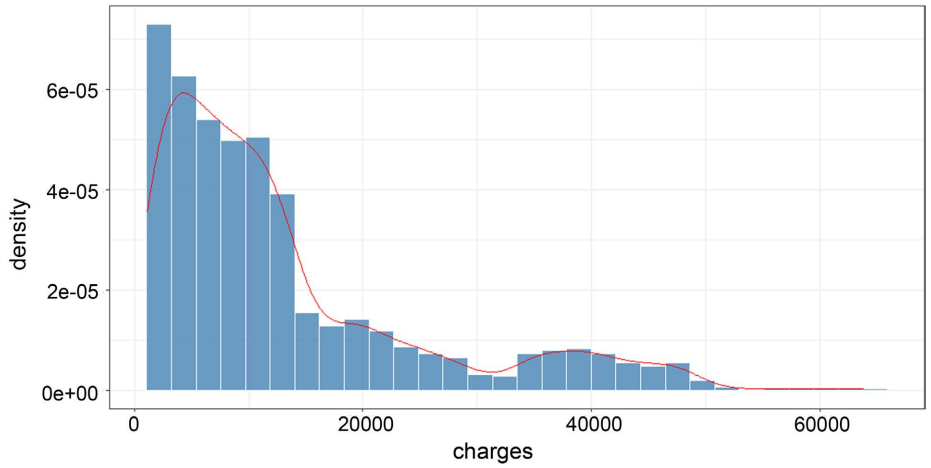

Figure 1. Charges distribution map
图 1. Charges(保险费)分布图

结合部分离散型变量，探究抽烟、性别、家庭方位中，不同水平的变量对保险费有怎样的影响。在图 2 中，可以看到不抽烟的人和抽烟的人之间平均保险费相差较大，且不抽烟的人平均保险费较低。不抽烟的人平均保险费在 10,000 元以下，且呈现右偏分布；而抽烟的人平均保险费在 30,000 元以上，且呈现双峰分布，从极差来看，抽烟的人保险费跨度大，不抽烟的人保险费较集中。

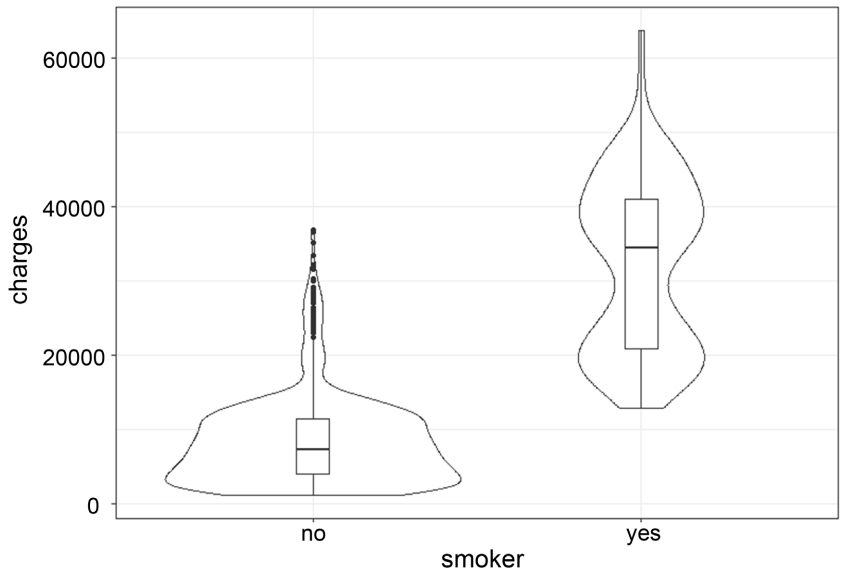

Figure 2. The effect of smokers on charges
图 2. Smoker (是否抽烟)对 charges (保险费)的影响

从图 3 中可以看到，不同方位的家庭保险费平均值均在 10000 元左右，且都呈现右偏分布，说明 region (家庭方位)不同对保险费没有明显的影响。

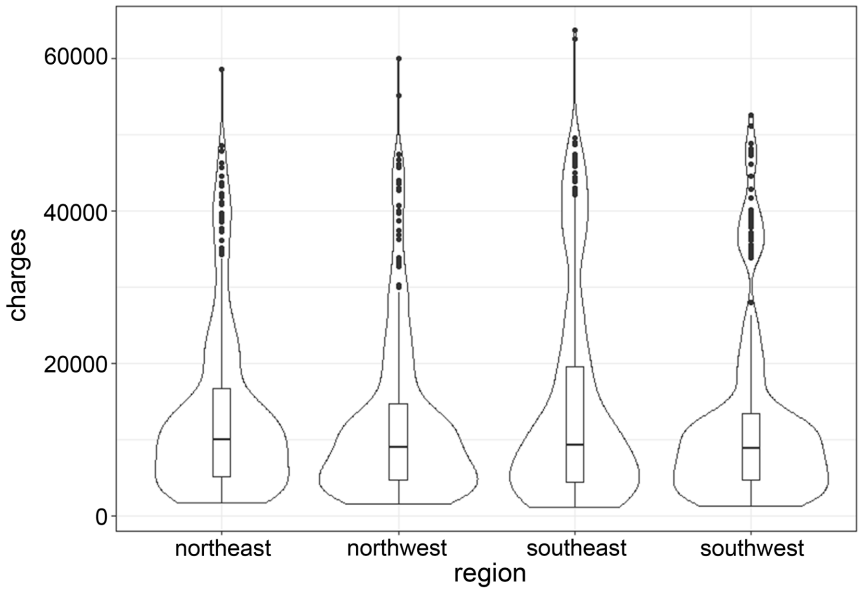

Figure 3. The impact of region on charges
图 3. Region (家庭方位)对 charges (保险费)的影响

图 4 中，男或女的保险费分布也几乎一致，且都呈现右偏分布，说明不同性别之间的保险费大致相同，性别不同不会导致保险费不同。

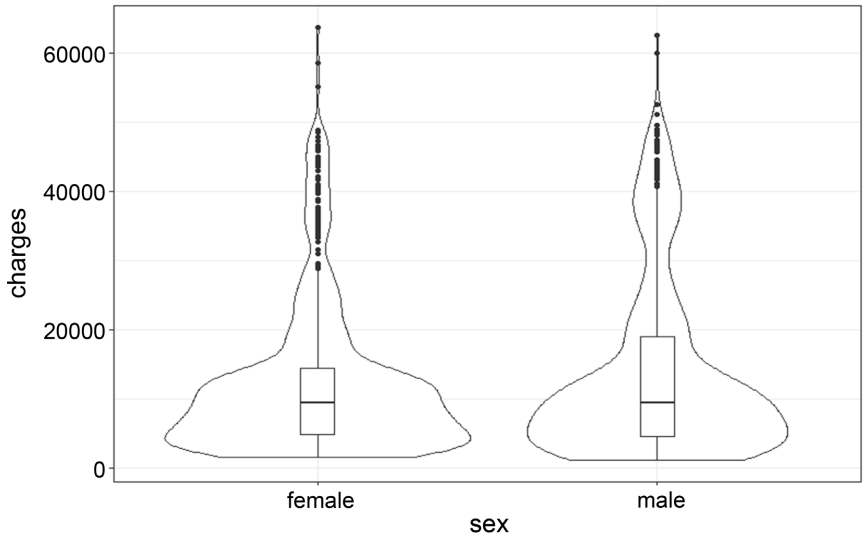

Figure 4. The impact of sex on charges
图 4. Sex (性别)对 charges (保险费)的影响

做 charges (保险费)与 age (年龄)、BMI (健康指数)、children (被保险家庭儿童数)的相关系数表，表 2 中，charges (保险费)与 age (年龄)的相关系数最高为 0.299，说明这三个因素都与 charges (保险费)的线性相关性较低，即保险费的提高和 age (年龄)、BMI (健康指数)、children (被保险家庭儿童数)的关系不大。

在图 5 中，不吸烟的消费者在各个年龄段的保险费都显著小于吸烟者在各个年龄段的消费者，各个年龄段的不吸烟者的保险费更集中，而各个年龄段吸烟者的保险费分布更加散乱。因此，是否吸烟可以认为是 charges (保险费)的一大重要影响因素。

Table 2. The correlation coefficient table
表 2. 相关系数表

	Children	BMI	Age
Charge	0.068	0.198	0.299

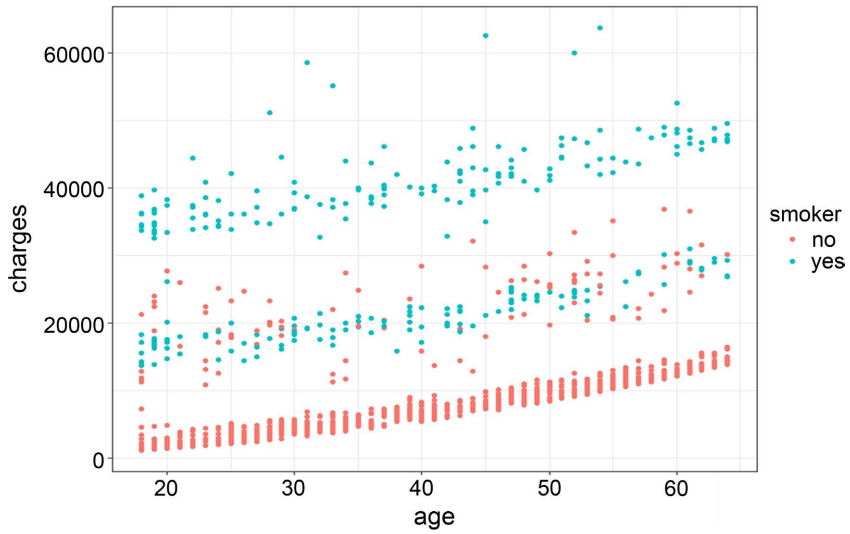

Figure 5. The impact of different age smokers on the charges
图 5. 不同 age (年龄)的吸烟者对 charges (保险费)的影响情况

在图 6 中，BMI (健康指数)的理想值为 18.5~24.9，在此区间内 charges (保险费)的波动较小，而在 BMI (健康指数)较高的区域，charges (保险费)整体波动的范围变大，且最大值提升较大，说明 BMI (健康指数)较为理想的人，charges (保险费)可能较低。

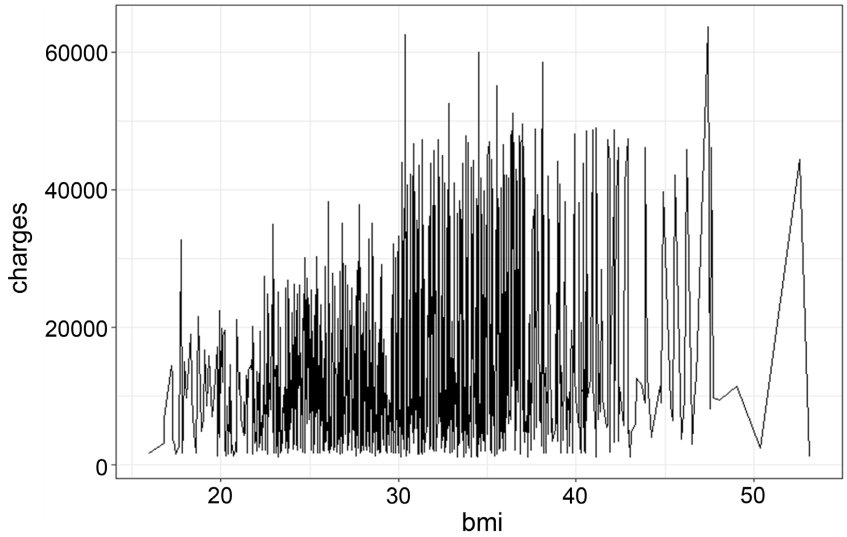

Figure 6. The impact of different BMI on charges
图 6. 不同 BMI (健康指数)对 charges (保险费)的影响

从以上分析来看，smoker (是否吸烟)对 charges (保险费)的影响最大，BMI (健康指数)也有部分影响，但和保险费的相关性不高。下面通过模型来衡量这些因素对 charges (保险费)的影响程度。

3. 随机森林预测分析

3.1. 随机森林介绍

随机森林是由 Leo Breiman 所创造的重要的机器学习方法[2]，随机森林的基本单元是决策树，决策树是随机森林模型的最重要也是最基础的要素之一，它通过每次随机抽取部分样本，对每个节点，在抽取部分特征进行拟合，构建出多棵决策树，它解决了单棵决策树存在的过拟合问题，且由于每次只选择部分变量，因此它又有类似聚类等筛选变量的作用。

从随机森林结果看，它集合了 n 棵决策树，对于一条样本，会产生 n 个结果，以最小均方误差或投票次数最多的结果为最终结果作为输出。

具体构建决策树算法如下：

1) 记 N 为所有样本数，M 为所有变量数。

2) 随机选取 m 个变量，m < M；n 个样本，n < N，确定决策树上节点最佳的分列方式，未抽到的样本作为训练集，计算误差。

3) 每棵树构建完不剪枝，再次构建下一棵树。

3.2. 指标介绍

由于决策树每次只选取部分样本拟合，剩余部分作为测试集，拟合完模型后，可以得到变量的重要性(variable importance)，它是每个特征在每棵树上贡献的平均值，体现了变量对模型的影响程度，如果变量的越高，说明替换该变量导致模型精准度下降的程度越大；拟合优度变量代表整个模型的精准度，拟合优度越高，说明模型能很好地解释因变量的变动[3]。

3.3. 模型结果分析

以 charges(保险费)为因变量，其余变量为自变量，随机挑选样本量的 80%作为训练集，剩余 20%样本作为测试集拟合随机森林模型。从模型来看，在训练集上，模型的拟合优度(var explained)为 84.26%，说明模型能解释 charges(保险费)变动的 84.26%。而在测试集，模型的均方误差为 14,067.79，模型效果较好。

在表 3 中，%IncMSE 表示从替换该变量而导致精确度平均递减的角度来衡量变量的重要性，IncNodePurity 则表示该变量为拆分变量所造成的均方误差的平均递减的角度来衡量变量的重要性，两个指标数字越大，说明变量越重要。

表 3 中，无论从%IncMSE 或是 IncNodePurity 来看，smoker (是否吸烟)对 charges (保险费)都有很强的影响，说明在考虑 charges (保险费)时，吸烟与否都是必须要关注的点；BMI (健康指数)的重要性排名第二，说明健康指数也是作为衡量身体健康程度、衡量 charges（保险费）的重要指标，age (年龄)的重要性排名第三，说明在考虑 charges (保险费)时，这几个因素不容忽视。

Table 3. A measure of the importance of random forests to individual variables
表 3. 随机森林对各个变量的重要性度量

	%IncMSE	IncNodePurity
Age	25,283,936.54	20,331,921,050
Sex	−80705.67	1,402,644,862
BMI	34,812,210.34	23,258,795,046

Continued

Children	3,670,095.37	3,751,485,793
Smoker	201,122,163.51	94,092,836,308
Region	1,173,281.79	2,732,890,504

将上表内容画在图 7，可以看到，sex (性别)、region (家庭方位)、children (被保险家庭儿童数)对模型的重要性不大，从%IncMSE 来看，sex (性别)甚至会使模型精确度下降，说明在考虑保险费时，不用考虑性别会对保险费产生较大的影响。

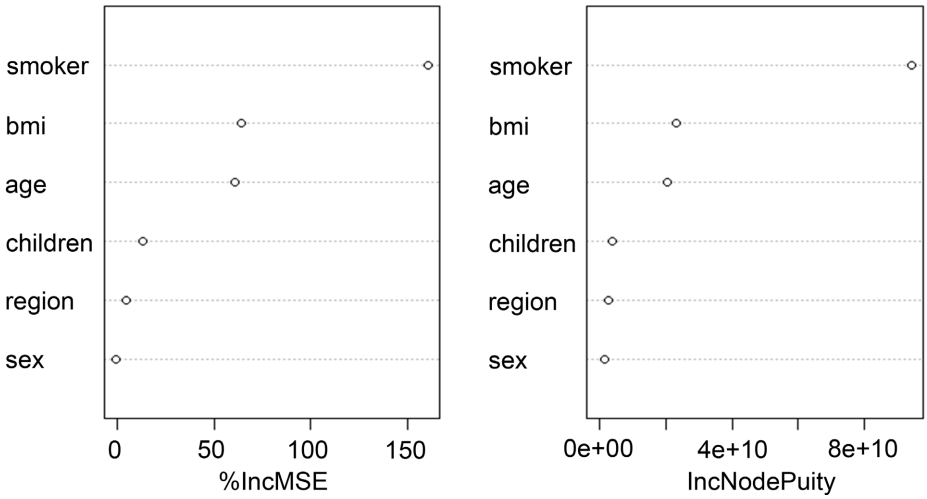

Figure 7. The diagram of the importance of the variable
图 7. 变量重要性图

4. 结论

本文通过探索性分析，发现 smoker (是否吸烟)在 charges (保险费)上的差别较大，不吸烟的人平均保险费较低，吸烟的人平均保险费较高，且吸烟的人保险费呈现双峰分布，说明如果吸烟，消费者购买保险所需缴纳的保险费在很大程度上会较高。对于这一现象分析结果说明吸烟者的健康会受到极大影响，会提高保险公司赔付的概率，因此保险费也越高。另外，BMI (健康指数)作为身体健康程度的测量指标，在考虑保险费时，也应当把它作为一个衡量指标。

从随机森林的结果来看，其拟合优度达到 84.26%，在变量重要性排序中，children (被保险家庭儿童数)、region (家庭方位)与 sex (性别)在模型当中都属于不重要变量，此结果与现实相符，说明模型拟合效果良好。

本文所拟合的模型效果良好，并且与实际相符，无论是消费者在针对自身情况购买保险或者是保险公司针对不同人群设置保险费用时，都可以使用这个模型进行分析。

参考文献

[1] 陆秋君, 陈玲, 施锡铨. 基于变结构协整理论的保费预测模型[J]. 数学的实践与认识, 2011(3): 1-7.

[2] 吴喜之. 复杂数据统计方法: 基于 R 的应用[M]. 北京: 中国人民大学出版社, 2012.

[3] 方匡南, 吴见彬, 朱建平, 谢邦昌. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.