

# 基于随机森林模型的船员选择培训机构影响因素分析

袁玉峰<sup>1,2,3,4</sup>, 郁晓红<sup>1,2,3,4</sup>, 郑彭军<sup>1,2,3,4\*</sup>

(1.宁波大学 海运学院, 浙江 宁波 315832;

2.浙江省 2011 港口经济协同创新中心, 浙江 宁波 315832;

3.国家道路交通安全管理工程技术研究中心宁波大学分中心, 浙江 宁波 315832;

4.现代城市交通技术江苏高校协同创新中心, 江苏 南京 210096)

**摘要:** 选取培训机构地理位置、培训机构许可培训项目数、培训机构服务质量以及船员年龄、籍贯、已参加培训次数作为影响因子, 采用随机森林模型对浙江省船员选择培训机构的影响因素进行了分析. 结果表明: (1)地理位置对船员培训机构选择的影响最大, 可解释 58.7%的船员选择培训机构的行为, 年龄对船员培训机构选择的影响最小; (2)影响因素的作用存在区域差异性, 同一影响因素对不同区域的船员选择培训机构的影响程度有所不同, 地理位置对舟山地区船员选择行为的影响最大, 对省内其他地区影响较小.

**关键词:** 船员选择培训机构行为; 影响因素; 区域差异; 随机森林

**中图分类号:** U6-9

**文献标志码:** A

**文章编号:** 1001-5132 (2021) 03-0079-05

船员是一种专业技能要求很高的职业, 船员适任资格的取得和维持都要求船员参加相应的培训, 船员通过海事主管部门的考试才能取得和保持相应的证书. 80%以上的海难事故都涉及人为因素<sup>[1]</sup>, 加强船员教育与培训是提高船员素质, 减少人为失误, 保护海上生命、财产安全和海洋环境的主要途径. 目前我国共有 234 家培训机构开展 56 项项目的船员培训, 2018 年培训量达 191 371 人次, 船员培训的重要性日益突出. 然而, 目前无论是培训机构的布局还是培训计划的制订都还没有建立统筹机制, 与船员对于培训质量和便利性的要求还存在较大差距<sup>[2]</sup>. 为了厘清影响船员选择行为的关键因素, 从而优化船员培训资源配置, 提高培训质量, 科学制订培训计划, 开展影响船员选择培训机构影响因素的研究十分必要.

目前, 有关船员培训的研究主要集中在培训过程中存在的问题以及如何提高船员培训质量两个方面, 而对于船员培训选择行为的影响因素鲜

有涉及. 在影响因素研究领域, 多位学者应用非集计模型<sup>[3]</sup>、多元回归模型<sup>[4]</sup>、决策树<sup>[5-6]</sup>、随机森林模型<sup>[7]</sup>、支持向量机模型<sup>[8]</sup>、K 近邻模型<sup>[9]</sup>、BP 神经网络模型<sup>[10]</sup>、Adaboost 模型<sup>[11]</sup>等方法开展了研究. 例如, 刘炳恩等<sup>[3]</sup>利用非集计离散选择模型结合 2003 年北京居民出行调查数据, 对影响居民出行方式选择的因素进行了分析, 并将出行者个人特性引入模型, 提高了模型的精度和实用性. Meng 等<sup>[4]</sup>利用多元回归分析了逆温、大气边界层高度、风向和相对湿度等不同气象因子对北京市冬季 PM<sub>2.5</sub> 浓度的相对贡献. 王文富<sup>[10]</sup>研究了影响企业人力资源结构的因素, 采用人工神经网络理论建立了预测模型. 柳本民等<sup>[12]</sup>以美国公路 2013~2015 年所有的追尾事故数据为样本, 建立了基于 SVM 的双车追尾事故与连环追尾事故二分类模型, 得到了导致连环追尾事故发生的关键影响因素.

应用机器学习进行影响因素的分析具有传统方法不具备的优势, 部分学者开展了一些有益的

收稿日期: 2020-09-15.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家重点研发计划政府间国际重点专项(2017YFE9134700); 欧盟地平线 2020 项目(690713); 浙江海事局科技项目(HS2019000170).

第一作者: 袁玉峰(1993-), 男, 安徽合肥人, 在读硕士研究生, 主要研究方向: 港航技术与管理工程. E-mail: 1593565982@qq.com

\*通信作者: 郑彭军(1965-), 男, 浙江慈溪人, 教授, 主要研究方向: 海事技术. E-mail: zhengpengjun@nbu.edu.cn

尝试,但也存在只关注机器学习模型的拟合精度,缺乏对模型本身的解释<sup>[13]</sup>等不足,而应用机器学习对船员选择培训机构影响因素分析尚未有报道.随机森林模型无需进行变量选择,可借助特征重要性排序方法进行因素分析,从而提高模型的可解释性<sup>[14]</sup>.因此,本研究采用随机森林模型,应用特征重要性排序增强模型的可解释性,来辨析各影响因素对选择行为的相对重要性.其次,利用模型验证集的拟合精度( $R^2$ )量化影响因素对船员选择培训机构的解释力度.

## 1 材料与方法

### 1.1 数据来源

选取2015~2019年浙江省6家船员培训机构船员培训考试的数据,其中杭州1家、宁波2家、舟山2家、温州1家,共73 894条记录.每条记录包含船员姓名、身份证号、培训机构、考试科目、成绩等内容.通过对原始样本数据整理分析,考虑到本文研究的对象为影响船员选择培训机构的因素,剔除只参加过1家机构培训的船员的记录,筛选出参加过1家以上机构培训的船员样本,得到23 521条样本数据.

### 1.2 特征变量选择

通过对现有研究的分析,影响选择行为的因素主要包括两个方面:(1)参与对象的个人属性,如年龄、性别、职业等;(2)被选择对象的社会属性,如地理位置、公众满意度、培训质量等.特别是年龄这一特征变量,在众多关于选择行为的研究中都显示了举足轻重的作用.如张荣花等<sup>[15]</sup>在研究汽车共享条件下居民选择出行方式的影响因素时,证明了年龄会对汽车共享服务选择行为产生显著影响;王慧芬等<sup>[16]</sup>在探究影响汽车共享换乘地铁选择行为的因素时,得到年龄、性别、职业等因素对汽车共享换乘地铁影响显著;练胜等<sup>[17]</sup>在研究居民的自行车选择行为影响因素时同样也得到了年龄是影响居民自行车选择行为的因素之一.因此,本文以船员访谈为基础,根据船员选择培训机构时会考虑的因素,并结合国内外相关学者的研究,选取船员年龄(AG)、所在地(HT)、参培次数(EXN)作为代表船员个人属性的特征变量以及培训机构的地理位置(GL)、许可培训项目数(TRN)、

服务质量(SQ)作为代表培训机构社会属性的特征变量,见表1.其中年龄和所在地从船员身份证号中提取,培训机构的服务质量根据培训机构的开班频率、培训人数、平均合格率及当地海事部门服务水平确定<sup>[18]</sup>.

表1 特征变量

| 变量     | 变量取值                                                        |
|--------|-------------------------------------------------------------|
| 年龄     | 20 以下=0, 20~29=1, 30~39=2, 40~49=3, 50~59=4, 60 及以上=5       |
| 所在地    | 本省=1, 外省=0                                                  |
| 参培次数   | $n$                                                         |
| 地理位置   | 本地市=1, 相邻地市=2, 其他=3                                         |
| 许可培训项目 | 根据海事局公布的各培训机构许可的培训项目数, 大于等于5项取1, 小于5项取0                     |
| 服务质量   | 服务水平高的为机构A和D, 取值为3; 服务水平中的为机构C和E, 取值为2; 服务水平较低的为机构B和F, 取值为1 |

### 1.3 方法

决策树算法是Quinlan<sup>[5]</sup>提出的,决策树模型中最常用的算法是基于信息熵的决策树学习算法——ID3 算法<sup>[6]</sup>,它的属性选择标准是信息增益,通常选择信息增益最大的属性作为它的“最佳”分裂点.其原理如下:设 $S$ 为一个包含 $s$ 个样本的集合,类别属性可以取 $m$ 个不同的值,对应于 $m$ 个不同的类别 $C_i$ ,  $i \in \{1, 2, 3, \dots, m\}$ .若属性 $A$ 选为测试属性, $A$ 有 $v$ 个不同的值 $\{a_1, a_2, a_3, \dots, a_v\}$ , $A$ 将集合 $S$ 划分为 $v$ 个子集 $\{S_1, S_2, \dots, S_v\}$ ,设 $N_{ij}$ 为子集 $N_j$ 中属于 $C_i$ 的样本数.划分当前样本集合所需要的信息可按下式计算:

$$E(A) = \sum_{i=1}^v \left( \frac{S_{1j} + S_{2j} + \dots + S_{mj}}{s} \right) \times I(S_{1j}, S_{2j}, \dots, S_{mj}), \quad (1)$$

对给定的子集 $s_j$ ,其信息为:

$$I(S_{1j}, S_{2j}, \dots, S_{mj}) = - \sum_{i=1}^m P_{ij} \log_2 P_{ij}, \quad (2)$$

式中, $P_{ij} = S_{ij} / |S_j|$ 为 $S_j$ 子集中样本属于类别 $C_i$ 的概率.则利用属性 $A$ 将当前的分支结点划分为相应样本集合后,再计算其信息增益:

$$G(A) = I(S_1, S_2, \dots, S_m) - E(A). \quad (3)$$

对每个属性的信息进行计算后,选择增益最大的属性作为给定集合 $S$ 的测试属性,并由此产生

相应的分支结点.

随机森林模型是 Breiman<sup>[7]</sup>于 2001 年提出的一种基于决策树的集成算法,它使用 bootstrap 重采样技术进行采样,从原始训练样本集中提取并生成训练样本子集,将提取到的训练样本子集生成决策树,由这些决策树组成的集合构成随机森林,最后取所有决策树的预测平均值作为最终预测结果.随机森林回归是由很多弱回归器(决策树)集成的强回归器.

在建立随机森林模型之前,需要先确立参数.从随机向量 $(X, Y)$ 中随机抽取  $n$  个训练集(通常训练集所占比例为 0.7),且这些随机向量均为独立同分布.其中  $X$  为输入向量,  $Y$  为输出向量.

设随机参数向量为  $\theta$ , 对应的决策树为  $T(\theta)$ . 记  $Q$  为  $X$  的域, 且  $Q \subseteq R^p$ , 其中  $P \subseteq N$  为自变量的维度. 决策树的每一个叶节点都对应一个矩形空间( $Q$ ), 记  $R_l$  且  $R_l \subseteq Q$ . 对任意  $x \in Q$ , 当且仅当一个叶节点  $l$  满足  $x \in R_l$ , 记决策树  $T(\theta)$  的叶节点为  $l(x, \theta)$ . 随机森林的建模过程如下<sup>[19]</sup>:

(1)利用 bootstrap 方法重采样, 随机产生  $k$  个训练集  $\theta_1, \theta_2, \dots, \theta_k$ ; 利用每个训练集生成对应的决策树  $\{T(x, \theta_1)\}, \{T(x, \theta_2)\}, \dots, \{T(x, \theta_k)\}$ .

(2)假设所有特征的维度为  $M$ , 从中随机抽取  $m$  个特征作为当前节点的分裂特征集合, 选择最佳的分裂方式对该节点进行分裂(通常,  $m$  的值在整个森林的生长过程中不会发生变化).

(3)随机森林中每个决策树都尽量以最大限度的方式增长, 不需要进行剪枝.

(4)在新的数据中, 通过叶节点  $l(x, \theta)$  的观测平均值可以得到单棵决策树  $T(\theta)$  的预测值. 假如一个观测值  $X_i$  属于叶节点  $l(x, \theta)$  且不为 0, 令权重  $\omega_i(x, \theta)$  为:

$$\omega_i(x, \theta) = \frac{1\{X_i \in R_l(x, \theta)\}}{\#\{j: X_j \in R_l(x, \theta)\}}, i=1, 2, \dots, n, \quad (4)$$

式中权重之和等于 1.

(5)根据因变量观测值  $Y_i$  ( $i=1, 2, \dots, n$ ) 的加权平均得到单棵决策树的预测. 单棵决策树的预测值的计算式为:

$$\mu(x) = \sum_{i=1}^n \omega_i(x, \theta) Y_i. \quad (5)$$

(6)通过对决策树权重  $\omega_i(x, \theta_i)$  ( $i=1, 2, \dots, k$ ) 取平均得到每个观测值  $Y_i \in (1, 2, \dots, n)$  的权重  $\omega_i(x)$ :

$$\omega_i(x) = \frac{1}{k} \sum_{i=1}^k \omega_i(x, \theta_i) Y_i, \quad (6)$$

则随机森林回归的预测值可记为:

$$\hat{\mu}(x) = \sum_{i=1}^n \omega_i(x) Y_i. \quad (7)$$

本文基于 23 521 个匹配样本数据, 将所有影响因素作为解释变量, 船员选择的培训机构作为因变量, 构建随机森林回归模型, 将训练集和测试集按 7:3 划分, 计算出模型精度( $R^2$ ), 并以此衡量影响因素对船员培训选择的影响大小.

## 2 结果分析

### 2.1 随机森林模型适用性评估

将影响因素作为自变量, 船员培训机构(机构 A~F 分别赋值 1~6)作为因变量, 构建多元线性回归模型、K 近邻回归模型、AdaBoost 回归模型和随机森林模型, 拟合结果见表 2. 随机森林模型的拟合精度( $R^2$ )最高, 误差(RMSE)和残差平方和(RSS)都最小, 因此, 选用随机森林模型分析船员培训选择行为的影响因素.

表 2 模型拟合结果

| 模型          | $R^2$   | MSE     | RMSE    | RSS         |
|-------------|---------|---------|---------|-------------|
| 多元回归        | 0.695 2 | 0.792 1 | 0.890 0 | 5 589.378 1 |
| K 近邻回归      | 0.784 6 | 0.561 8 | 0.749 5 | 3 963.760 0 |
| AdaBoost 回归 | 0.748 8 | 0.645 3 | 0.803 3 | 4 552.938 6 |
| 随机森林        | 0.841 3 | 0.411 1 | 0.641 1 | 2 900.428 6 |

### 2.2 影响因素重要性分析

本文提出的随机森林模型可以量化单个因素对船员选择的影响程度. 分别将各影响因素作为输入变量, 以船员选择的机构值作为输出变量, 训练随机森林模型, 模型拟合结果如图 1 斜对角线所示: 按影响程度大小排序, 地理位置、服务质量、许可培训项目数、参培次数、所在地和年龄分别能解释 58.7%、43.8%、4.1%、1.0%、0.5%和 0.1%的船员培训选择行为. 地理位置和服务质量对船员培训选择行为的影响程度最高.

同理, 影响因素两两组合作为随机森林模型的输入变量, 可以量化双因素交互作用对船员培训选择行为的影响, 从图 1 可见, 地理位置与服务质量交互能够解释 81.2%的船员培训选择行为, 地理位置与许可培训项目数能解释 64.7%的船员培训选择行为.

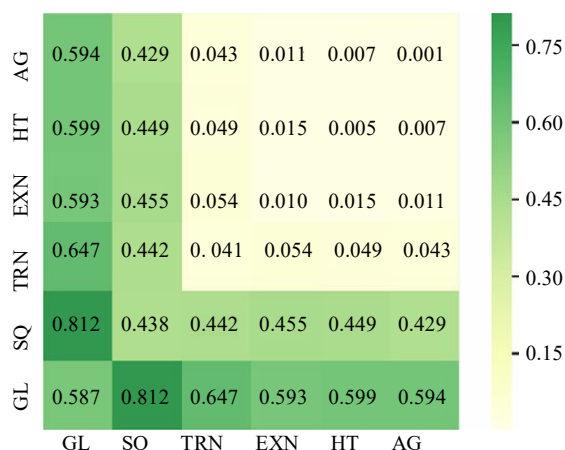


图1 船员选择培训机构各因素的影响程度

### 2.3 影响因素的区域差异分析

为进一步探索船员选择行为影响因素的区域化差异,按船员所在地,分别对沿海地区船员人数较多的舟山、宁波、台州、温州的船员数据进行了分析,结果如图2所示。随机森林模型的拟合精度对所有地区船员均较高( $R^2>0.85$ )。

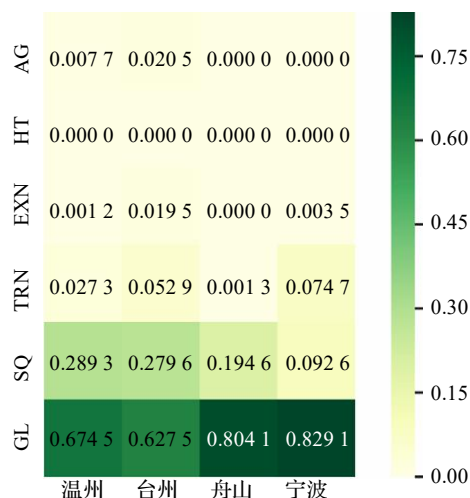


图2 不同地区船员选择培训机构的影响因素的解力度

从结果可以看出,影响不同地区船员选择行为的主要因素排序十分相似,宁波、舟山和温州、台州均为地理位置和服务质量,而且培训机构的地理位置影响远远大于其他因素,说明船员有就近选择培训机构的特点。

对于宁波和舟山的船员,地理位置可以解释80.4%和82.9%的选择;对于台州和温州的船员,服务质量可以解释28.9%和28.0%的船员选择。船员培训选择行为的最主要因素地理位置的影响存在一定区域差异,其解释率在不同地理分区的排

序为:宁波>舟山>温州>台州。这主要是因为宁波和舟山的船员培训机构规模较大,许可培训项目较齐全,本地船员基本可以选择就地培训。温州地区虽然有一家培训机构,但规模很小,培训项目有限(仅有2项培训资质),而台州没有本地培训机构,因此大多数培训需要选择其他地区的培训机构,考虑了较多的培训质量因素,地理位置的解释力度就相对减少了。不同地区的年龄、籍贯以及考试次数对船员培训选择的地区差异性解释力度均很小,说明船员选择行为的区域差异化与这三者的相关性不显著。

### 3 结论

(1)地理位置对船员培训机构选择行为的影响最大,可解释58.7%的船员培训机构选择行为;年龄对船员培训机构选择行为的影响最小,解释率仅为0.1%。因素交互显著提高了影响因素对船员培训机构选择行为的解释力度。

(2)船员培训机构选择行为与影响因素间的关系存在地区差异性,同一影响因素对不同地区的船员培训机构选择行为的影响力度存在区域差异。地理位置对宁波、舟山地区船员培训机构选择行为的影响较大;温州、台州地区由于缺少本地培训机构,服务质量对本地区船员选择行为的影响较大。

(3)影响船员选择培训机构的最主要因素是培训机构的地理位置和服务质量。如果有本地培训机构,大多数船员会优先选择本地培训机构,服务质量的影响会相对较小,说明合理布局培训机构对船员培训服务非常重要。在选择本地机构受限时,服务质量对船员选择培训机构的影响会增加,因此,培训机构也可通过提高培训服务质量来扩大服务范围,增强培训机构的影响力。

### 参考文献:

- [1] 关腾飞. 论述船员培训存在的问题及建议[J]. 珠江水运, 2003(8):25-26.
- [2] 邓慧云. 高等教育资源共享问题与对策研究[D]. 开封: 河南师范大学, 2016.
- [3] 刘炳恩, 雋志才, 李艳玲, 等. 居民出行方式选择非集计模型的建立[J]. 公路交通科技, 2008(5):116-120.
- [4] Meng C, Cheng T H, Gu X F, et al. Contribution of meteorological factors to particulate pollution during

- winters in Beijing[J]. *Science of the Total Environment*, 2019, 656:977-985.
- [5] Quinlan J R. Induction of decision trees[J]. *Machine Learning*, 1986, 1(1):81-106.
- [6] Quinlan J R. C4.5: Programs for Machine Learning[M]. San Diego: Morgan Kaufmann, 1993.
- [7] Breiman L. Random forests[J]. *Machine Learning*, 2001, 45(1):5-32.
- [8] 丁良萍, 张智雄, 刘欢. 影响支持向量机模型语步自动识别效果的因素研究[J]. *数据分析与知识发现*, 2019, 3(11):16-23.
- [9] 尹晓梅, 李梓铭, 熊亚军, 等. 2014-2017北京市气象条件和人为排放变化对空气质量改善的贡献评估[J]. *环境科学*, 2019, 40(3):1011-1023.
- [10] 王文富. 企业人力资源结构的神经网络预测[J]. *天津大学学报(社会科学版)*, 2005, 7(1):20-22.
- [11] 王坤, 刘小杰, 刘二浩, 等. 基于AdaBoost算法的炉芯温度预测模型[J]. *钢铁研究学报*, 2020, 32(5):363-369.
- [12] 柳本民, 闫寒. 基于 SVM 事故分类的连环追尾事故影响因素分析[J]. *交通信息与安全*, 2020(1):43-51.
- [13] Li J M, Jin M J, Li H L. Exploring spatial influence of remotely sensed PM<sub>2.5</sub> concentration using a developed deep convolutional neural network model[J]. *International Journal of Environmental Research and Public Health*, 2019, 16(3):E454.
- [14] 王超, 阚媛珂, 曾业隆, 等. 基于随机森林模型的西藏人口分布格局及影响因素[J]. *地理学报*, 2019, 74(4):664-680.
- [15] 张荣花, 关宏志, 赵磊, 等. 考虑汽车共享服务的出行方式选择影响因素研究[J]. *交通运输研究*, 2020, 6(3):48-54.
- [16] 王慧芬, 王立晓, 孙小慧. 基于 Logit 模型的汽车共享换乘地铁选择行为研究[J]. *交通科技与经济*, 2020, 22(4):33-38.
- [17] 练胜, 李晓达, 黄舒晴, 等. 基于分层 Logit 模型的自行车选择行为影响因素调研[J]. *科技与创新*, 2017(23):8-10.
- [18] 中华人民共和国海事局. 中华人民共和国海事局关于公布 2019 年 1-11 月海船船员培训机构适任考试及格率的公告[EB/OL]. [2019-12-31]. <http://www.msa.gov.cn/public/documents/document/mtax/nduy/~edisp/20191231101452703.wps>.
- [19] 方匡南. 随机森林组合预测理论及其在金融中的应用[M]. 厦门: 厦门大学出版社, 2012.

## Analysis of factors influencing ship crew's choice of training institutions based on random forest model

YUAN Yufeng<sup>1,2,3,4</sup>, YU Xiaohong<sup>1,2,3,4</sup>, ZHENG Pengjun<sup>1,2,3,4\*</sup>

(1.Faculty of Maritime and Transportation, Ningbo University, Ningbo 315832, China;

2.Zhejiang Province 2011 Port Economic Collaborative Innovation Center, Ningbo 315832, China;

3.National Traffic Management Engineering & Technology Research Centre Ningbo University Sub-centre, Ningbo University, Ningbo 315832, China;

4.Jiangsu Province Collaborative Innovation Center for Modern Urban Traffic Technologies, Nanjing 210096, China )

**Abstract:** The geological location of the training institution, the number of training projects licensed by the training authorities, the quality of training services, the age of the ship crew, the place of birth of the crew, and the number of trainings having participated by the crew have been selected as the influencing factors in this study. The random forest model is used to analyze the factors influencing the selection of training institutions by crew members in Zhejiang Province. The results show that: (1) Geographic location has the most significant impact on selecting the training institutions, which may account for 58.7% of the crew's choice of training institutions; in comparison, age has the least influence. (2) The role of influencing factors varies according to regions. Also noted is that the same influencing factors have different impacts on crew's selection of training institutions in different regions. Geographic location has the greatest impact on the crew's selection preference in Zhoushan area, and less is found in other areas across the province.

**Key words:** crew's selection of training institutions; influencing factor; regional differences; random forest

(责任编辑 韩超)