Modeling and Simulation 建模与仿真, 2023, 12(6), 5531-5542
Published Online November 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.126502

Hans汉斯

基于机器学习的员工流失预测

李思晴，罗鄂湘

上海理工大学管理学院，上海

收稿日期：2023年8月20日；录用日期：2023年11月17日；发布日期：2023年11月23日

摘　要

员工流失是当今组织中的重要问题，通过机器学习等技术对员工离职进行事前预测，有助于提升企业人力资源管理的前瞻性。本文首先从采集的数据集中提取有用的且适合模型训练条件的数据，进行数据清洗和探索性分析，了解各特征分布情况；使用One-Hot编码和标签编码相结合的方式进行编码，然后采用RF-RFE方法对数据集中的特征进行筛选后进入分类模型；为保证模型预测的准确性，采用了五种不同的机器学习算法，包括SVM、DT、RF、LightGBM和WRF，来建立模型对员工流失情况进行预测。综合结果显示，LightGBM算法在预测性能方面表现出色，其准确率达到了0.87；进而通过SHAP输出特征重要性排名，发现城市发展指数、工作经验和培训时数等是影响离职的重要因素，可以为企业员工保留和后续人才招聘决策提供技术支持。

关键词

员工流失预测，机器学习，RF-RFE，LightGBM，WRF

Employee Attrition Prediction Based on Machine Learning

Siqing Li, Exiang Luo

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Aug. 20th, 2023; accepted: Nov. 17th, 2023; published: Nov. 23rd, 2023

Abstract

Employee turnover is an important issue in today's organizations. Predicting employee turnover in advance through machine learning and other technologies can help improve the foresight of enterprise human resource management. In this paper, the useful data are extracted from the collected data set, and the data cleaning and exploratory analysis are carried out to understand the distribution of features. A combination of One-Hot coding and label coding was used, and then the features in the dataset were screened by RF-RFE method and entered into the classification model.

文章引用: 李思晴, 罗鄂湘. 基于机器学习的员工流失预测[J]. 建模与仿真, 2023, 12(6): 5531-5542.
DOI: 10.12677/mos.2023.126502

In order to ensure the accuracy of the model prediction, five different machine learning algorithms, including SVM, DT, RF, LightGBM and WRF, were used to build the model to predict the employee turnover situation. The results show that LightGBM algorithm has excellent prediction performance, and its accuracy rate reaches 0.87. By outputting SHAP feature importance ranking, it is found that urban development index, work experience, training hours, etc., are important factors affecting turnover, which can provide technical support for employee retention and follow-up talent recruitment decisions.

Keywords

Employee Attrition Prediction, Machine Learning, RF-RFE, LightGBM, WRF

1. 引言

员工流动是“人员在组织内进入和退出就业的流动”[1]。在相对集中的时期内大量离职，会给管理带来困难。从组织的角度来看，不必要的员工流失会给组织带来相当大的负担：正在进行的项目的中断、员工再培训的成本、核心技术泄漏的风险等。与过去事后再来处理员工流失不同，现在有可能通过人工智能提前预测员工流失的可能性来采取先发制人的行动，用以辅助支持下一阶段的人才招聘决策过程。

近年来，机器学习等数据分析技术的应用作为一种预防员工流失的方法而备受关注。根据德勤咨询发布的《2017 全球人力资本趋势人力资源分析：路线重新规划》，显示在参与的各个公司中，高达 72%的公司认为数据在公司的人力资源发展中扮演着至关重要的角色。数据挖掘技术的应用能够极大地提升企业的人力资源管理效率。特别是在员工流失率分析方面，利用数据挖掘技术能够有效地解析不同岗位员工离职的原因、比例以及对企业带来的损失等关键信息。

在员工流失预测方面，Ebru Pekel Ozmen 等(2022) [2]构建一种混合深度学习算法应用于零售行业的员工流失预测的数据集，为预测员工流失提供了一种有效的方法；刘春燕(2021) [3]基于随机森林和 XGBoost 构建了员工流失预测模型，在模型评估中发现 XGBoost 的预测表现更佳。对与衡量员工离职的诸多因素进行筛选时，王冠鹏(2022) [4]使用了由 Cui 等(2015) [5]提出的 MV (Mean of Variance)方法和卡方法相结合进行降维，并与 LASSO 方法降维进行了对比，提升了变量筛选的稳健性。而对于不平衡的数据，万毅斌等(2022) [6]应用了改进的代价敏感加权 SVM 算法进行分类预测，取得了良好效果。

本文建立支持向量机(Support Vector Machines, SVM)、决策树(Decision Tree, DT)、随机森林(Random Forest, RF)、轻量化梯度促进机(Light Gradient Boosting Machine, LightGBM)以及加权随机森林(Weighted Random Forest, WRF)多个分类模型，并进行了详细的对比分析。研究首先对部分特征进行了可视化分析，然后采用 RF-RFE (Random Forest-Recursive Feature Elimination)方法对数据进行了特征筛选，提取了影响员工流失的关键特征。随后，基于这些关键特征构建了员工流失预测模型，采用 SVM、DT、RF、LightGBM 和 WRF 算法，以有效分类离职员工等少数类样本。

2. 理论基础

2.1. 扩展 Mobley 模型

William H. Mobley 等[7]于 1979 年结合多种模型后提出了 Mobley (1979)模型，为解释员工流失现象

提供了一个理论框架，模型中将这些离职因素分为三大类，即个体因素、组织因素和外部因素。其中个体因素就包括员工的性别、受教育水平以及个人特点和态度等；组织因素包括组织的规模、类型以及政策等；外部环境因素包括社会、经济、行业趋势以及地理位置等，外部环境因素可能对员工流失产生间接影响，影响就业机会、竞争和工作市场。

本文将从个体、组织和外部环境因素这三个方面深入探讨各因素对员工流失的影响，以揭示员工流失的多维度机制。同时，将有助于组织制定更有效的员工留存策略，减少流失率。

2.2. 算法介绍

在机器学习中，常见的预测算法包括支持向量机、决策树、随机森林等。支持向量机是一种经典的二分类模型，其核心原理是寻找一个最佳分类超平面，以确保分类准确率的同时使分类间隔最大化，通常用于回归和分类问题[8]；决策树是一种基于树状结构的监督学习算法，主要特点是其可解释性和易于理解，但容易在训练数据上过拟合[9]；随机森林是一种监督学习算法，对异常值和噪声具有良好容忍度，不容易过拟合，但当决策树数量较多时，模型训练需要更多的时间和空间资源。除此之外，本文还引入了近年来在预测方面表现较为优秀的算法：LightGBM 和 WRF。

这些机器学习算法在不同情况下都具有独特的优势，可以根据具体问题和数据集的特点选择适合的方法来构建高性能的预测模型，在此对 LightGBM 和 WRF 做重点介绍。

2.2.1. LightGBM 算法

LightGBM 的本质是一种将弱学习器提升为强学习器的集成学习算法，通过单边梯度采样算法来最大限度地保留对计算信息增益有帮助的样本，加快模型的训练速度。方差增益为

$$V_j(d)=\frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_{l,1}}G_i+\frac{1-a}{c}\sum_{x_i\in A_{l,2}}G_i\right)^2}{n_l^j(d)}+\frac{\left(\sum_{x_i\in A_{r,1}}G_i+\frac{1-a}{c}\sum_{x_i\in A_{r,2}}G_i\right)^2}{n_r^j(d)}\right)$$

其中：n 为样本总数，j 为使用的分裂特征，d 为样本特征的分裂点，$n_l^j(d)$、$n_r^j(d)$ 分别为使用的分裂特征值小于和大于 d 的样本数，$A_{l,1}$、$A_{l,2}$ 分别为所分裂的左子节点中的大、小梯度样本，$A_{r,1}$、$A_{r,2}$ 分别为所分裂的右子节点中的大、小梯度样本，G 为样本梯度，a 为大梯度样本采样率，c 为小梯度样本采样率。

LightGBM 的特点和优势在于：1) 高效性：采用直方图增强技术，将连续特征分桶成直方图，以提高训练速度和降低内存占用，适用于大规模数据集。2) 按叶子结点生长：采用按叶子结点生长的策略，减少过拟合风险，更快地找到有信息量的叶子结点。3) 并行训练支持：支持多线程和分布式计算，提高训练效率。4) 高度优化的损失函数：提供多种优化的损失函数，提高模型性能。5) 特征选择和重要性评估：自动计算特征重要性，帮助特征选择和模型优化。

2.2.2. 加权随机森林算法

通常 RF 的所有决策树在投票进行分类时都具有相同的权重值。但实际情况下，一些分类精度低的决策树很可能会投出错误的票数，并对最终的分类结果造成影响。为此，本文将加权 F1-measure 引入 RF 算法，通过为不同的决策树分配不同的权重，为员工流动率预测提供更好的性能。

$$\omega_t=\frac{X_{correctt}}{X}\qquad t=1,2,\cdots,T$$

其中：ω_t 为第 $X_{correctt}$ 棵树正确分类的样本数；X 为预测试样本数；决策树数量为 T。将此正确率作为对应决策树的权重，则加权后随机森林模型的输出为

$$\hat{F} = \arg\max\left\{\sum_{f_t(x)=i} \omega_t\right\} \quad i = 1, 2, \cdots, p$$

其中：ω_t 为第 t 棵决策树的权重值。如图 1 所示，为加权随机森林的结构图。运用 WRF 算法构建员工流失率的预测模型，是将全部样本分成训练集和验证集，用训练集创建多个基分类器，每个基分类器用验证集进行预测并计算 f1 值作为权重，创建 WRF 算法。

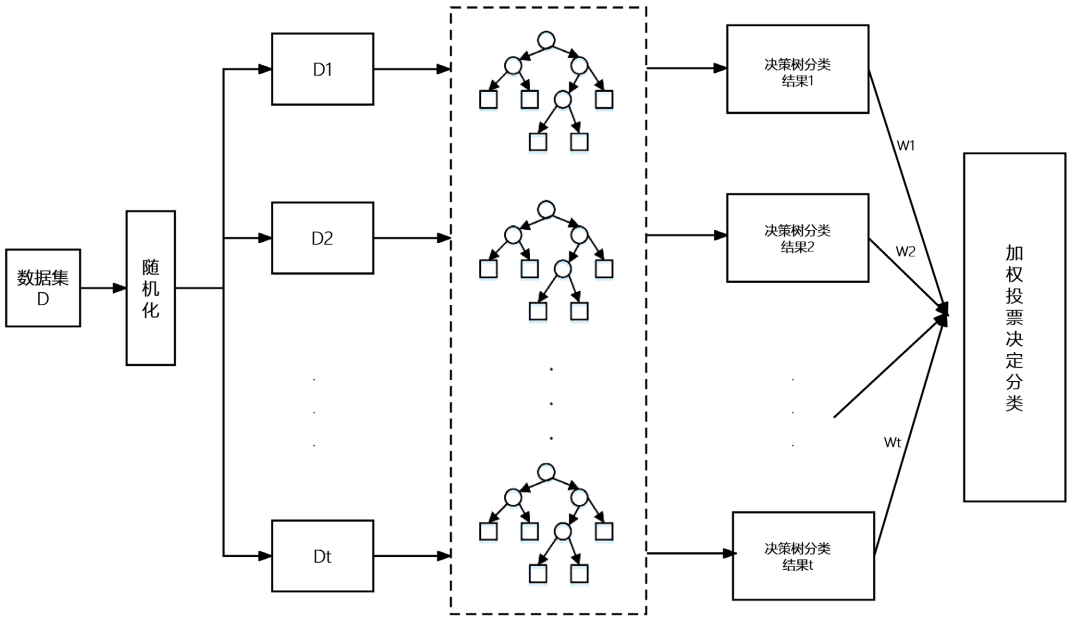

Figure 1. Structure of weighted random forest
图 1. 加权随机森林结构图

2.3. 技术路线

本文首先收集数据，从中提取有用的且适合模型训练条件的数据；其次进行数据清洗，剔除一些无关变量和相似变量，以及对重复值和缺失值进行处理；然后进行特征选择，基于全部样本采用 RF-RFE 方法并应用交叉验证(Cross Validation, CV)进行特征筛选，选出对于员工流动较为重要的特征 m；随后，用这些 m 特征代入到 SVM、DT、RF、LightGBM 和 WRF 算法中，建立模型；最后预测模型可以通过准确率、精确率、召回率、f1 score 和 AUC 进行评估，选出最佳模型。

以下是运用各类算法进行员工流失预测的技术路线图，如图 2 所示。

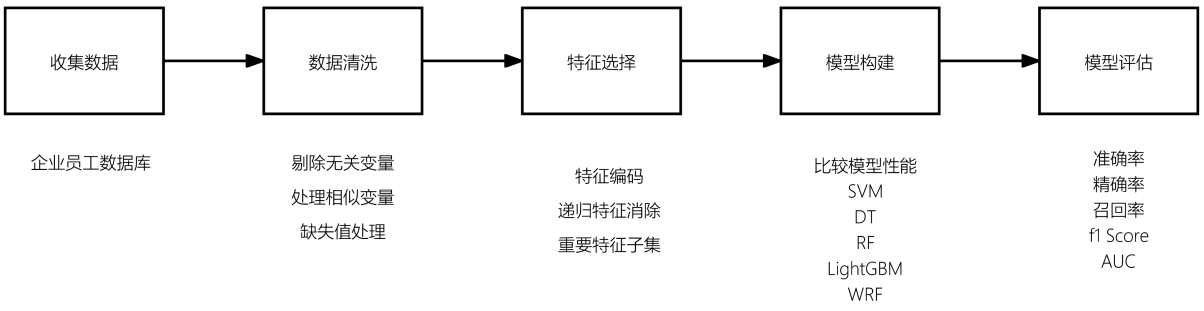

Figure 2. Technology roadmap
图 2. 技术路线图

3. 数据说明及处理

3.1. 数据描述与清洗

3.1.1. 数据描述

本文使用 Kaggle 平台 2020 年提供的用于人力资源分析的数据集，该数据集由一家主要从事大数据和数据科学相关业务的公司提供，数据集中包括了参加该公司开设的培训课程的人员信息，共有 19,158 条数据，包括 13 个特征和一个目标变量“是否离职”(target)，目标变量中的“0”表示“没有在寻找工作机会”，“1”表示“正在寻找工作机会”。

对此数据集，本文首先提取了只含有主修科学、技术、工程和数学教育(STEM)方向的人员数据，因为原始数据集中 STEM 专业的人数占比约为 76%，且 STEM 人才对社会和经济的重要性不可忽视，他们不仅推动了科技进步和经济增长，还有助于解决全球性问题和提高国家的全球竞争力，所以我们这里仅对 STEM 的员工进行实验研究，以发现 STEM 人才流失的原因，为公司保留 STEM 人才提出有效建议。在此基础上提取了“gender”(性别)中填写了“male”和“female”的人员信息。

因此本文研究的数据基本情况如表 1 所示，包括 13 个变量和 11,073 个样本。

Table 1. Basic information of features
表 1. 数据基本情况

变量	名称	变量类别	值域
enrollee_id	员工工号	名义变量	[1, 33380]
city	员工所在城市	名义变量	[city_1, city_103]
city_development_index	城市发展指数	连续变量	[0.448, 0.949]
gender	性别	二分类变量	Female, Male
relevent_experience	相关经验	二分类变量	No relevant experience, Has relevant experience
enrolled_university	大学类型	定序变量	no_enrollment, Part time course, Full time course, nan
education_level	教育水平	定序变量	Graduate, High School, Master, Phd, Primary School, nan
experience	工作经验(年)	连续变量	[0, 21]
company_size	公司规模	定序变量	<10, 10~49, 50~99, 100~499, 500-999, 1000~4999, 10,000+, nan
company_type	公司类型	名义变量	Early Stage Startup, Funded Startup, NGO, Other, Pvt Ltd, Public Sector
Last_new_job	现在工作与上一份工作间隔年数	连续变量	1, 2, 3, 4, >4, never, nan
training_hours	完成培训所用的时间	连续变量	[1, 336]
target	是否会流失	二分类变量	0-Not looking for job change 1-Looking for a job change

3.1.2. 数据清洗

在现实生活中，采集的原始数据就不可避免的存在一些问题，被称之为“脏数据”，会对模型的准确性产生一定的影响，因而需要通过数据清洗去填补一些残缺的数据，纠正错误的数据，统一数据的格

式等。本文对数据主要采用了以下的清洗方法：

① 剔除无关变量。在原始数据集中，包含“enrollee_id”这样一个表示“员工工号”的特征，对于预测模型无实际意义，予以删除。

② 处理相似变量。考虑到数据集中变量“city”与变量“city_development_index”所表达意思相近，且对于目标变量的影响也相似，为避免重复分析，因而将“city”这个特征作了删除处理。

③ 考虑到变量“company_type”中的“Other”样本量只有 57 个，并且不明确其具体的公司类型，缺乏实质性的研究意义，因此决定将相关样本删除。

④ 重复值和缺失值处理。首先删除重复值，然后将存在缺失的数据进行剔除。

数据清理后，剔除后的数据集包含 12 个变量 7831 条样本，此数据集为非平衡数据，样本中目标变量为“1-Looking for a job change”的占比仅为 17.3%。

3.2. 探索性数据分析(EDA)

根据扩展 Mobley 离职模型，本文将从个体因素、企业因素和外部因素三个方面来探讨各因素对于员工流失的影响。以下是对各类因素的详细描述，并对有关因素进行可视化。

3.2.1. 个体因素

个体因素中的 EDA 分析主要做了 Gender (性别)、Education Level (教育水平)、Last New Job (最近一次换工作距今时间)、Training Hours (培训时数)，如图 3 所示。

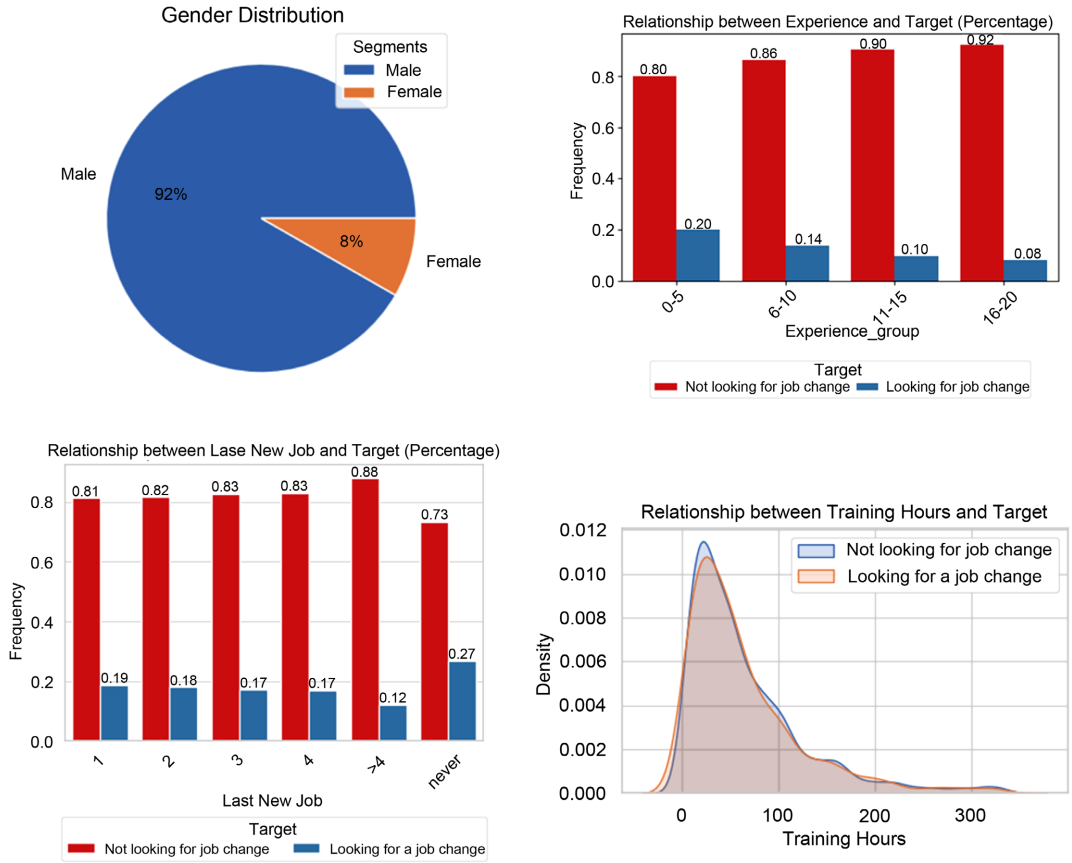

Figure 3. Individual factors
图 3. 个体因素

从图 3 中首先可以看出，此数据集中以男性员工为主，占总人数的 92%，女性仅占 8%，这可能与 STEM 的专业有关；其次可以发现，Experience (工作经验)柱状图中显示出，随着成员工作经验的增加，离职率逐步降低，在一定程度上说明，成员工作经验的年数越多，工作的稳定性就越高；然后通过 Last New Job (最近一次换工作距今时间)柱状图中可以看出，距离最近一次换工作间隔四年以上的样本离职率相对较低，可见长期内未换工作的员工对于工作的忠诚度相对更高；最后从 Training Hours (培训时数)密度图中分析可以看出，培训时长呈右偏态分布，以相对较短时间的培训为主，有意向流动的人员与无意向流动人员的样本在培训时间的分布上没有明显区别。

3.2.2. 组织因素

组织因素中主要包括 Company Size (公司规模)和 Company Type (公司类型)两方面，对其进行可视化，如图 4 所示。

Figure 4. Organizational factors
图 4. 组织因素

从图 4 中就可以看出不同规模的公司员工的流失情况，样本中流失率较高的是在 10~49 人规模的小型公司工作的员工，占比为 22%。另外，不同类型的公司，如初创公司、公共部门、私营公司等，就会有不同的流失率，Early Stage Startup (初创公司)的离职率相对较高。

3.2.3. 外部环境因素

本数据集中的外部环境因素主要是 City Development Index (城市发展指数)，对其绘制密度图，如图 5 所示。从图 5 中可以观察到，没有工作变动意愿的样本主要分布在发展指数约为 0.9 左右的城市，有工作变动意愿的样本所在城市的发展指数分布则相对分散。这可能表明发展迅速的城市因其综合实力较强，吸引了大量人才，员工也更愿意留在这些城市，不愿意轻易更换单位。

3.3. 基于随机森林的特征选择

3.3.1. 数据编码

在本研究中，对不同类型的特征进行了合适的数值编码，以便用于机器学习模型的训练和分析。本文的编码处理如表 2 所示，编码后，自变量个数由 11 个变为 14 个，样本数为 7831 个。对于连续型变量“city_development_index”、“experience”以及“training_hours”则不需要进行编码。

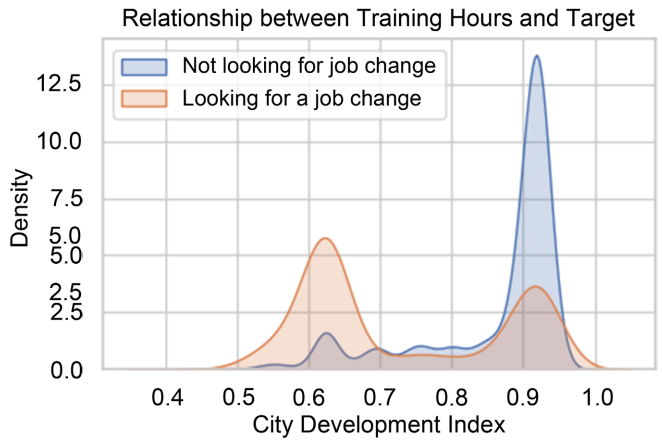

Figure 5. External environment factor
图 5. 外部环境因素

Table 2. Feature coding correspondence
表 2. 特征编码情况

	变量	赋值
1	gender	male = 0, female = 1
2	relevent_experience	No relevant experience = 0, Has relevant experience = 1
3	enrolled_university	no_enrollment = 0, Part time course = 1, Full time course = 2
4	education_level	Graduate = 0, Master = 1, Phd = 2
5	company_size	<10 = 0, 10~49 = 1, 50~99 = 2, 100~499 = 3, 500~999 = 4, 1000~4999 = 5, 10,000+ = 6
6	last_new_job	never = 0, 1, 2, 3, 4, >4 = 5
7	company_type	用 One-Hot 变成虚变量
8	target	Not looking for job change = 0, Looking for a job change = 1

3.3.2. 特征筛选

特征选择就可以从原始数据集中剔除掉部分非关键特征，得到影响目标变量的最优特征子集。本文采用了一种基于随机森林算法的递归特征消除(RF-RFE)的方法。

随机森林(Random Forest, RF)是一种由 Breiman 提出的组合算法。它的工作方式包括以下步骤：从全样本数据集 D 中，使用自举重采样方法随机抽取子样本 Dt，并利用这些子样本来训练 T 棵决策树。每棵决策树会对不同特征进行评分，最后通过投票方式来决定将样本分到哪一类。总之，随机森林通过集成多个决策树的结果，利用自举重采样和基尼系数来进行特征选择和数据划分，提供了一种强大的分类和回归方法，适用于各种数据集和问题。它的优点在于降低了过拟合的风险，提高了模型的鲁棒性和性能。方法为

$$\Delta Gini = Gini(D) - Gini_A(D)$$

其中 $Gini(D) = 1 - \sum_i^c p_i^2$，$c$ 表示不同的类别，P_i 表示类别 i 占整体比例的大小，即数据越混乱，相应 $Gini$ 系数值就越大。$Gini_A(D)$为选取的属性 A，分裂后数据集 D 的系数值，计算公式为

$$Gini_A(D) = \sum_j^k \frac{|D_j|}{|D|} Gini(D_j)$$

RF-RFE 结合了随机森林和递归特征消除的思想，旨在帮助机器学习从原始特征集中识别和选择最重要的特征，以提高模型的性能和减少过拟合的风险。本文在特征选择过程中，对清理后的非平衡数据和 SMOTE 后的平衡数据，分别运用了 RF-RFE 进行特征选择，结果如图 6 所示。为得到准确结果，本文分别求出交叉验证分割后的 5 个子集准确率的平均值进行验证，最终得到最佳特征数量为 10 个。

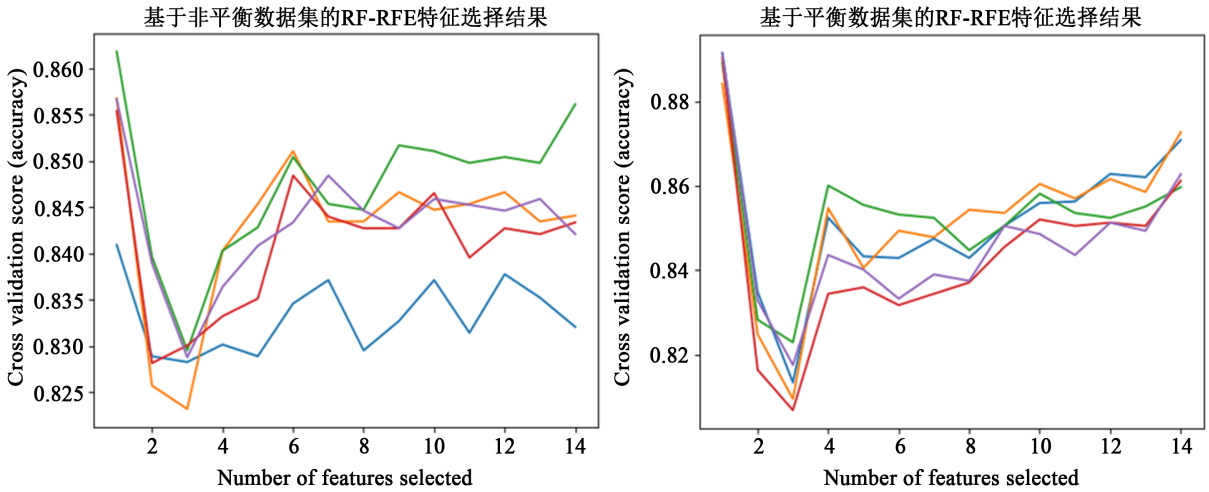

Figure 6. The result of feature selection
图 6. 特征选择结果

因此，本文在后续的建模过程中，选择 10 个变量进入模型，删去了排名较低的四个变量，分别是公司类型中的虚变量 Funded Startup (资助初创企业)、Public Sector (政府企事业)、Early Stage Startup (初期创业公司)和 NGO (非政府组织)。

4. 建模与评估

4.1. 模型建立与分析

本文模型建立及评估过程如下：首先，将 10 个变量和 7831 个样本的非平衡数据集和 SMOTE 之后的数据集均按 6:4 划分为训练集和测试集；其次，分别将非平衡数据和 SMOTE 之后的数据中的训练集代入到 SVM、DT、RF、LightGBM 和 WRF 算法中，建立模型；最后，用测试集计算各模型的准确率、精确率、召回率和 f1 score，并应用 SMOTE 之后的测试集数据绘制各模型 ROC 曲线。

结果如表 3 及图 7 所示。

Table 3. The results of the model of test set
表 3. 测试集模型结果

	模型	准确率	精确率	召回率	f1 score
均衡前	SVM	0.83	0.41	0.50	0.45
	DT	0.77	0.60	0.61	0.61
	RF	0.84	0.72	0.65	0.68
	LightGBM	0.85	0.73	0.67	0.69
	WRF	0.85	0.72	0.66	0.68

Continued

	SVM	0.80	0.80	0.80	0.80
	DT	0.84	0.84	0.84	0.84
均衡后	RF	0.86	0.86	0.86	0.86
	LightGBM	0.87	0.87	0.87	0.87
	WRF	0.86	0.86	0.86	0.86

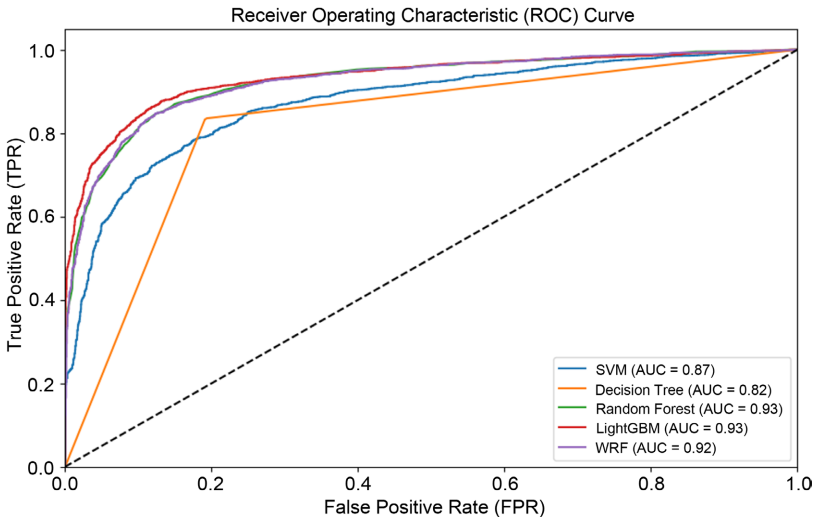

Figure 7. ROC curves of various algorithms
图 7. 各类算法的 ROC 曲线

根据表 3 的结果，可以看出，对于非平衡数据，WRF 算法和 LightGBM 算法在精确率、准确率上表现要优于 SVM、DT、RF 的性能，而在召回率和 f1 值上其他算法要略逊于 LightGBM 算法；对于 SMOTE 之后的平衡数据，LightGBM 的表现则优于其他算法。此外，从图 7 中观察到 LightGBM 算法的 ROC 曲线更接近左上角，这说明 LightGBM 在员工流失预测任务中具有更高的性能。

综上所述，LightGBM 算法在多个性能指标上都展现出了较为出色的预测能力，这表明它是一种有效的方法，可用于准确地预测员工流失情况，为企业提供更好的决策支持。

4.2. 变量重要性分析

在建立离职预测模型后，本文选用 SHAP (SHapley Additive exPlanations)来衡量某一特征对预测的影响程度大小，寻找对其离职与否影响较大的因素，从而采取合适的措施避免员工的流失。SHAP 是 Python 开发的一个“模型解释”包，基于合作博弈理论中的 shapley 值，来解释每个特征对于模型预测的贡献程度。如图 8 所示，SHAP 摘要图中揭示了各个特征如何正负反馈作用于目标变量，图中每个颜色点代表一个样本，颜色代表各员工流失变量的特征值的大小(红色高，蓝色低)。从图 8 中可以看出，城市发展指数至各公司类型的 SHAP 绝对值总和从上至下依次减小，对目标变量影响较大的变量“city_development_index”的 SHAP 值主要集中在负值且为红色，说明存在发展指数高的城市，其离职率相对较低。对员工流失有较大影响的因素还有：Experience (工作经验)、Training Hours (培训时数)、Last New Job (最近一次换工作距今时间)等。同时也可以看出 Company Type (公司类型)里的几个虚变量对于模型预测的贡献程度都相对较低，从而验证了前面特征筛选过程具有的稳健性。

Figure 8. Summary of SHAP attribution analysis
图 8. SHAP 归因分析摘要图

5. 结论

基于以上分析，本文可得到如下结论。

1) 针对影响员工流失的因素，通过 RF-RFE 方法进行变量选择，结果发现其中 10 个变量对目标变量有影响，结合 SHAP 摘要图进行特征重要性分析，发现以下特征的员工是否流失的影响较大。

个体因素方面：主要包括员工的性别、教育水平、最近一次换工作距今时间、培训时数。员工的工作经验年数与流失率之间存在着重要关联，具有更多工作经验的员工往往更加职业稳定。因此，企业在人事决策时需要考虑员工年资，这将有助于降低员工流失率，建议企业加强对新员工的培训，以培养他们对所从事职业的热情，从而提高员工粘性，减少离职率。

企业因素方面：公司规模和公司类型对员工是否有工作变动意愿产生了影响。小型公司和初创公司的员工流失率相对较高，规模较大的公司通常提供更具挑战性的晋升机会，但相对更加稳定。因此，小型公司要更关注员工的个性化需求，满足员工的个性化需求。

外部环境因素方面：城市的发展指数对员工流失率也有一定影响。城市的发展水平高通常意味着更多的就业机会和较高的职业稳定性。相比之下，发展水平较低的城市可能资源相对稀缺，可供选择的职位有限，从而导致员工更容易流失。但需要指出的是，较低发展水平的城市中的员工通常承受的工作压力相对较低。因此，建议在发展水平较低的城市中的企业在公司建设时优化员工的晋升途径，并提供更舒适的工作环境，如健身房、茶水间和休息室等，以提高员工满意度，减少流失率，增强员工的职业稳定性。

2) 在模型选择方面，本文通过建立 SVM、DT、RF、LightGBM 以及 WRF 等多个分类模型进行详细的分析和综合效果对比，发现 LightGBM 算法在综合指标上要优于其他算法，该方法的核心在于其高效的梯度提升算法和一系列优化技术，可以自动计算特征的重要性，帮助用户识别哪些特征对模型性能有最大贡献，从而进行特征选择和模型优化。在实际应用价值上，通过更准确地员工流失预测，企业可以采取有针对性的措施来改善员工满意度、留任率，从而提高组织的稳定性和竞争力。

参考文献

[1] Ann, D. and McMahon, F. (1992) Labour Turnover in London Hotels and the Cost Effectiveness of Preventative Measures. *International Journal of Hospitality Management*, **11**, 143-154. https://doi.org/10.1016/0278-4319(92)90007-I

[2] Ozmen, E.P. and Ozcan, T. (2022) A Novel Deep Learning Model Based on Convolutional Neural Networks for Employee Churn Prediction. *Journal of Forecasting*, **41**, 539-550. https://doi.org/10.1002/for.2827

[3] 刘春燕. 基于 XGBoost 的员工流失预测研究——以 IBM 公司为例[D]: [硕士学位论文]. 大连: 大连理工大学, 2021.

[4] 王冠鹏, 秦双燕, 崔恒建. 员工流失的影响因素分析与预测[J]. 系统科学与数学, 2022, 42(6): 1616-1632.

[5] Cui, H.J., Li, R.Z. and Zhong, W. (2015) Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association*, **110**, 630-641. https://doi.org/10.1080/01621459.2014.920256

[6] 万毅斌, 王绍宇, 秦彦霞. 基于代价敏感加权支持向量机的员工离职分类预测[J]. 智能计算机与应用, 2021, 11(12): 43-46+53.

[7] Mobley, W.H. (1977) Intermediate Link Ages in the Relationship between Job Satisfaction and Employee Turnover. *Journal of Applied Psychology*, **62**, 237-240. https://doi.org/10.1037/0021-9010.62.2.237

[8] 李芸, 胡可, 董欣雨, 袁淑俊. 基于 SVM 算法的企业员工离职预警研究[J]. 中国商论, 2020(6): 20-22.

[9] Abellán, J., Mantas, C.J., Castellano, J.G. and Moral-García, S. (2018) Increasing Diversity in Random Forest Learning Algorithm via Imprecise Probabilities. *Expert Systems with Applications*, **97**, 228-243. https://doi.org/10.1016/j.eswa.2017.12.029

[10] Gomes, H.M., Bifet, A., Read, J., Barddal, J.P., Enembreck, F., Pfharinger, B., Holmes, G. and Abdessalem, T. (2017) Adaptive Random Forests for Evolving Data Stream Classification. *Machine Learning*, **106**, 1469-1495. https://doi.org/10.1007/s10994-017-5642-8