Advances in Applied Mathematics 应用数学进展, 2022, 11(3), 1420-1427
Published Online March 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.113155

Hans汉斯

基于Logistic回归及机器学习方法对IBM员工流失因素的实证分析

常沐冉[1]，黄坷杰[1]，张元巨[2]，王　志[1]

[1]宁波工程学院，理学院，浙江　宁波
[2]宁波技师学院，基础部，浙江　宁波

收稿日期：2022年2月28日；录用日期：2022年3月22日；发布日期：2022年3月29日

摘　要

通过建立Logistic回归模型和机器学习模型对IBM员工的基本特征进行分析。基于员工的年龄、部门、受教育程度等特征值，处理相关数据，设置虚拟变量，进一步构建Logistic回归模型及机器学习模型，实证分析各类特征值对员工流失率是否有显著影响，并找出最优拟合及预测效果。同时，从这些相关因素出发，针对不同属性的员工提出相应的解决方法，降低员工的流失率。最后，对不同方法的模型进行比较，为模型的可靠性提供参考标准。

关键词

Logistic回归模型，机器学习，员工流失

Prediction and Comparative Analysis of IBM Employee Turnover Based on Logistic Regression and Machine Learning

Muran Chang[1], Kejie Huang[1], Yuanju Zhang[2], Zhi Wang[1]

[1]School of Sciences, Ningbo University of Technology, Ningbo Zhejiang
[2]Department of Basic, Ningbo Technician College, Ningbo Zhejiang

Received: Feb. 28th, 2022; accepted: Mar. 22nd, 2022; published: Mar. 29th, 2022

Abstract

The basic characteristics of IBM employees are analyzed by Logistic regression model and ma-

文章引用: 常沐冉, 黄坷杰, 张元巨, 王志. 基于 Logistic 回归及机器学习方法对 IBM 员工流失因素的实证分析[J]. 应用数学进展, 2022, 11(3): 1420-1427. DOI: 10.12677/aam.2022.113155

chine learning model. Based on the employee's age, department, education level and other characteristic values, the logistic regression model and machine learning model are constructed to analyze and predict whether the employee loses. At the same time, it analyzes the correlation between various factors and employee turnover rate, and puts forward corresponding solutions based on these factors, so as to ensure that employees have a sense of belonging, improve their working conditions and reduce the employee turnover rate. Finally, different models are compared to select the most appropriate model to provide a more accurate reference for predicting the turnover of employees.

Keywords

Logistic Regression Model, Machine Learning, Staff Turnover

1. 引言

知识经济迅猛发展的今天，人力资源日益成为企业获得竞争优势的重要源泉，而员工流失却是当前困扰许多企业的重要问题之一。2019 年我国的员工离职率达到 18.9%，其中主动离职 13.4%，对任何一家企业，人力资源管理都至关重要，而就我国的现状而言，在劳动力上一直处于供不应求的状态，劳动力需求较大[1]。员工的流失可能意味着大量的行业信息或科技成果的流失，许多的产品甚至一片市场被带走，亦或原来的经营计划受阻，商业秘密的泄漏，其他员工的积极性的挫败等。员工流失问题已经逐渐为中国企业管理人士所关注。影响员工流失的个人因素主要有年龄、工作年限、受教育程度、家庭状况及生活方式等因素[2]。

文章使用 Logistic 回归模型和机器学习方法，分析各种因素与员工流失率之间的相关性，并从这些因素出发，对员工离职原因进行汇总分析。通过分析及时发现公司管理中存在的问题，并了解员工的真实心理状态，为公司改进管理和政策调整提供参考，对发现的问题及时整改，不断提高员工满意度的凝聚力，保证公司快速稳定发展。

当今各大企业的人员流动较大，新老员工的交替十分频繁，如何保留员工，增加员工的归属感，让企业拥有更多优秀的人才，对于一家企业来说是非常重要的，具有较强的现实意义，也有利于企业与员工的互利共赢[3] [4] [5]。

2. 数据的收集及预处理

本文的数据来源于 IBM 数据科学家创建的数据集，包含员工的年龄、部门、受教育程度等信息，一共 1470 行，35 个字段。主要研究影响员工流失率的因素，所以将员工是否流失作为因变量，该变量是一个二分类变量，1 表示没有流失，0 表示流失。因此预测该变量的情况时，考虑采用分类方法进行预测。

自变量为剔除因变量员工是否流失后的所有特征变量，查看自变量数据无缺失值，考虑到自变量对因变量的影响情况，去除一些无关数据。在数据预处理中，其中有 4 个属性：员工数量、员工编号、是否超过 18 岁、标准时间四个变量无法为研究提供有效信息，因此剔除这 4 个变量，剩余 31 个字段。

3. Logistic 回归模型的分析及预测

Logistic 回归模型是一种广义线性回归，回归模型中，y 是一个定性变量，比如 $y=0$ 或 1，Logistic

方法解决的是分类问题，主要应用于研究某些事件发生的概率[6]。

3.1. 模型的建立及变量选择

1) 模型的建立

Logistic 回归模型是一种研究二项分类结果与多个互不相关的影响因子$(x_1, x_2, \cdots, x_n)$之间关系的多元统计分析方法。本文主要研究影响员工流失率的因素，其中自变量为用户基本特征因子指标值$(x_1, x_2, \cdots, x_n)$，因变量为被调查者是否离职，分别为 1 和 0。自变量可为连续变量、离散变量或二者任意组合，无需满足正态分布。构建 Logistic 回归模型如下：

$$P(y=1)=\frac{1}{1+e-(\alpha+\beta_1 x_1+L+\beta_n x_n)} \tag{1}$$

式中：P 为用户离职的概率；α 为截距；β 为回归系数。P 的输出结果是 0~1，为了便于理解与计算，对上式左右两端取自然对数，得线性回归方程如下：

$$\ln\left(\frac{p}{1-p}\right)=\alpha+\beta_1 x_1+L+\beta_n x_n \tag{2}$$

2) 变量的选择

将所有数据进行汇总，纳入所有需要考虑的变量，设置随机种子，按照 8:2 比例划分训练集和测试集。建立多元逻辑回归 Logistic 模型，对训练集进行拟合并估计参数，通过逐步回归，经过数次变量筛选，最终得到 AIC 值最小的模型，其各显著变量的系数估计值、标准误差和对应 P 值，结果如下表 1 所示：

Table 1. Table of model parameter estimation results
表 1. 模型参数估计结果表

特征	系数估计值	标准误差	P 值
截距	−3.1978	0.18595	2.00E−16
年龄	−0.4307	0.12536	0.000591
商务差旅频率	0.3983	0.10092	7.94E−05
日工作效率	−0.2294	0.09862	0.019993
所在部门	−0.3682	0.10475	0.00044
离家距离	0.4112	0.09568	1.72E−05
环境满意度	−0.4327	0.0997	1.42E−05
工作投入度	−0.3488	0.0969	0.000319
工作角色	0.2442	0.11087	0.027649
工作满意度	−0.5920	0.10103	4.65E−09
婚姻状况	−0.6068	0.10761	1.71E−08
月收入	−0.9082	0.1874	1.26E−06
曾在几家公司工作	0.4540	0.10228	9.05E−06
关系满意度	−0.2614	0.09947	0.008581
去年培训时间	−0.2577	0.10594	0.015007
工作与生活平衡程度	−0.2593	0.09656	0.007249

Continued

在目前公司时长	0.7244	0.24144	0.002695
在当前职位时长	−0.5576	0.18441	0.002495
距离上次升职时长	0.4219	0.15249	0.00566
和目前管理者共事年数	−0.6408	0.17885	0.00034
是否加班	2.0058	0.21032	2.00E−16

3.2. 模型的评估及预测

由上表 1 可知，回归方程中的显著性变量一共有 20 个，最后保留在模型中的这些解释变量既是重要的，又没有严重多重共线性。Logistic 回归方程如下：

$$\begin{aligned}\log it(p) = &-3.1978 - 0.4307x_1 + 0.3983x_2 - 0.2294x_3 - 0.3682x_4 - 0.3009x_5 \\ &- 0.4388x_6 + 0.5172x_7 + 0.2442x_8 - 0.5920x_9 - 0.6068x_{10} - 0.9082x_{11} \\ &+ 0.4540x_{12} + 2.0058x_{13} - 0.2614x_{14} - 0.2577x_{15} - 0.2593x_{16} + 0.7244x_{17} \\ &- 0.5576x_{18} + 0.4219x_{19} - 0.6408x_{20}\end{aligned}$$

通过得到的逻辑回归方程对测试集进行预测，得到模型评估结果如下表 2：

Table 2. Logistic model fitting accuracy index
表 2. Logistic 模型拟合精度指标

评价指标	精度值/%
准确率	0.8853
灵敏度	0.8853
特异性	0.68
测试集正确率	0.8684

由表 2 可知，Logistic 回归模型的预测准确率约为 87%，准确率约为 89%，说明模型判断为正样本的置信概率较高，该模型判断出的正样本可信；灵敏度约为 96%，说明分类器对正例的识别能力非常好，而特异度为 68%，相较而言，分类器对负例的识别能力不是很好，但是可以接受。

4. 机器学习模型的分析及预测

机器学习算法最初多用于解决回归问题，本次研究主要使用了监督分类问题中的算法[7] [8]。

4.1. 特征值的处理及选择

特征选择是为了解决高维度数据计算问题而衍生的，通过剔除冗余特征和无关特征，来提高机器学习算法的泛化性能和运行效率。本文由于数据有 object 类型，所以将非数值型特征重新进行独热编码、标签化等操作之后，通过对特征值进行相关性处理，画出与“是否离职”这一变量的热图，选取相关系数较高的特征值结果如下表 3 所示。

由此可知，这 12 个特征值与预测列“是否离职”相关性较强，所以选取这些特征值进行模型拟合以得到更好的结果。然后从数据源中随机抽取 80%的数据作为训练集，其余的为测试集，最后使用 StandardScaler 对特征值进行归一化处理，去均值和方差归一化，消除特征间单位和尺度差异的影响。

Table 3. Table of characteristic variable selection results
表 3. 特征变量选择结果表

特征变量	相关系数
年龄	−0.16
商务差旅频率	0.13
环境满意度	−0.10
工作投入度	−0.13
职业级别	−0.17
工作满意度	−0.10
婚姻状况	−0.16
月收入	−0.16
是否加班	0.25
在目前公司时长	−0.13
在当前职位时长	−0.16
和目前管理者共事年数	−0.16

4.2. 各类模型的建立及评估

4.2.1. K 近邻分类模型

KNN (K- Nearest Neighbor)即 K 最邻近法，最初由 Cover 和 Hart 于 1968 年提出，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。通过 sklearn.neighbors 模块建立 KNN 模型，对训练集数据进行 KNN 分类算法模型训练，并选择最优参数 n_neighbors 为 2。运用模型对测试数据进行预测，得到模型评估结果如下表 4：

Table 4. KNN model fitting accuracy index
表 4. KNN 模型拟合精度指标

评价指标	精度值%
准确率	0.881
灵敏度	0.984
特异性	0.211
测试集正确率	0.8707

由以上拟合及预测，得到该模型的正确率为 87.07%、准确率为 88.1%、灵敏度为 98.4%、特异性为 21.1%、AUC 值为 0.632。可以看出运用 KNN 模型进行拟合效果良好。

4.2.2. 决策树模型

决策树可看作一个树状预测模型，它通过把实例从根节点排列到某个叶子节点来分类实例，叶子节点即为实例所属的分类。通过 sklearn.tree 模块建立决策树模型，对训练集数据进行决策树分类算法模型训练，并选择最优参数 max_depth 为 5。运用模型对测试数据进行预测，得到模型评估结果如下表 5：

Table 5. Decision tree model fitting accuracy indicators
表 5. 决策树模型拟合精度指标

评价指标	精度值/%
准确率	0.883
灵敏度	0.973
特异性	0.132
测试集正确率	0.8639

由以上拟合及预测，得到该模型的正确率为 86.39%、准确率为 88.3%、灵敏度为 97.3%、特异性为 13.2%、AUC 值为 0.737。可以看出运用决策树模型进行拟合效果良好。

4.2.3. 随机森林模型

随机森林由 LeoBreiman (2001)提出，它通过自助法(bootstrap)重采样技术，从原始训练样本集 N 中有放回地重复随机抽取 k 个样本生成新的训练样本集合，然后根据自助样本集生成 k 个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的分数而定。通过 sklearn.ensemble 模块建立决策树模型，对训练集数据进行决策树分类算法模型训练，并选择最优参数 n_estimators 为 300。运用模型对测试数据进行预测，得到模型评估结果如下表 6：

Table 6. Fit accuracy index of stochastic forest model
表 6. 随机森林模型拟合精度指标

评价指标	精度值/%
准确率	0.907
灵敏度	0.996
特异性	0.31
测试集正确率	0.9014

由以上拟合及预测，得到该模型的正确率为 90.14%、准确率为 90.7%、灵敏度为 99.6%、特异性为 31%、AUC 值为 0.785。可以看出运用随机森林模型进行拟合效果良好。

4.2.4. SVM 模型

支持向量机(support vector machines)是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机。SVM 的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM 的学习算法就是求解凸二次规划的最优化算法。通过 sklearn.svm 模块建立决策树模型，对训练集数据进行决策树分类算法模型训练，并选择最优参数 C 为 10。运用模型对测试数据进行预测，得到模型评估结果如下表 7：

Table 7. SVM model fitting accuracy index
表 7. SVM 模型拟合精度指标

评价指标	精度值/%
准确率	0.871
灵敏度	1
特异性	0
测试集正确率	0.8707

由以上拟合及预测，得到该模型的正确率为 87.07%、准确率为 87.1%、灵敏度为 100%、特异性为 0%、AUC 值为 0.621。可以看出运用随机森林模型进行拟合效果还好。

5. 不同模型的对比及结论

通过上述多种方法进行模型的建立、训练和预测，我们根据模型预测的准确率、正确率、AUC 值三个评价模型优劣的指标对五种模型进行比较，得到下表 8：

Table 8. Fitting accuracy of different models
表 8. 不同模型拟合精度

模型	测试集正确率/%	准确率/%	AUC
多元 Logistic 回归	0.8684	0.885	0.658
KNN	0.8707	0.881	0.632
决策树	0.8639	0.883	0.737
随机森林	0.9014	0.907	0.785
支持向量机	0.8707	0.871	0.621

1) 从总体上来看，各个模型都能很好的拟合和预测本次研究项目，但最优模型为随机森林。

2) 随机森林模型拟合效果最好，其次是决策树和多元 logistics 回归，而 KNN 和支持向量机模型效果较差。

3) 随机森林和决策树对处理标签型和数值型数据有优势且分类精确度高，又因为随机森林是通过对模型进行重采样的方式，构建出多棵决策树，然后利用多棵决策树进行预测，那多棵树进行平均后，可以降低掉采样造成的偏差，所以随机森林比决策树预测效果更好。

4) 由于该样本特征空间较大，所以逻辑回归的性能相比与 KNN 的性能较弱。

6. 建议

本文通过对数据集的探索，发现了影响员工流失因素的相关特点及一套合适的预测员工是否流失的模型，根据研究我们可以为企业提供以下建议：

1) 建立和完善用人，育人和留人机制。企业应采取相应的对策和措施，尽快改革人力资源管理体制，努力建立公开、平等、竞争、择优的用人机制，完善对人才的激励和考核监督机制，加快建立有利于留住人才和人尽其才的收入分配制度，与国际接轨，确保人才工作朝着正规化、科学化的方向发展。

2) 完善企业文化。企业的核心竞争力主要是企业文化，是企业凝聚力和活力的源泉，企业文化是留住员工的关键因素。企业文化应该是包括企业的核心理念、经营哲学、管理方式、用人机制、行为准则、企业氛围的总和，是一个综合体。

3) 建立科学的薪酬绩效体系。首先要区分出不同岗位的工资等级，形成合理的纵向差别；其次要合理确定薪资水平，确定每个岗位的绝对工资数；通过定期对本地区、同行业进行薪资调查获取有关信息，可以制定出外部竞争性强而且又符合公司的支付能力的工资水平。

4) 建立员工流失预测预警管理系统，指对企业人才各种状况进行识别、分析、判断，并做出警示和调控的管理活动。建立员工流失预测管理系统，关键是在认真分析员工离职原因的基础上确定相应的预警指标，如工作满意度、工作压力感、员工对公司认同感等。预警系统建立后，实施对这些指标的日常监控，一旦它们偏离正常安全范围，系统立即发出预警信号，随后危机管理小组马上展开分析调查，弄清问题出现的原因并评估其影响程度，然后在此基础上做出相应的管理决策。

基金项目

宁波市自然科学基金(2019A610041)；国家级大学生创新创业项目(202111058033)；作者对王宽诚教育基金会的资助谨致谢忱。

参考文献

[1] 张萌. 国有企业员工流失的原因和对策分析[J]. 中小企业管理与科技(中旬刊), 2021(8): 96-98.

[2] 刘阴莺子. 基于员工满意度的企业员工流失原因分析[J]. 中国管理信息化, 2021, 24(3): 148-149.

[3] 张月寒. 企业核心员工流失影响因素分析及对策研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2007.

[4] 裴琳. 重庆市民营企业核心员工流失影响因素调查分析及对策研究[D]: [硕士学位论文]. 重庆: 重庆大学, 2006.

[5] 马金贵, 张长元. 企业核心员工流失原因分析及其对策[J]. 湖南商学院学报, 2005, 12(2): 42-44.

[6] 于立勇, 詹捷辉. 基于 Logistic 回归分析的违约概率预测研究[J]. 财经研究, 2004, 30(9): 15-23.

[7] 李旭然, 丁晓红. 机器学习的五大类别及其主要算法综述[J]. 软件导刊, 2019, 18(7): 4-9.

[8] 杨剑锋, 乔佩蕊, 李永梅, 王宁. 机器学习分类问题及算法研究综述[J]. 统计与决策, 2019, 35(6): 36-40.