Advances in Applied Mathematics 应用数学进展, 2024, 13(4), 1338-1344
Published Online April 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.134124

Hans汉斯

基于逻辑回归的衡阳市高职毕业生离职问题研究

李凌云，申娇娣，李合军，彭姣丽，孙　兴，骆又麟，陈　恒

湖南环境生物职业技术学院生态宜居学院，湖南 衡阳

收稿日期：2024年3月17日；录用日期：2024年4月11日；发布日期：2024年4月18日

摘　要

近年来，我国每年的应届高职毕业生在半年内发生离职的比例超过40%，这一现象应当引起各方广泛关注。本文向2022届衡阳市高职毕业生发放调查问卷，基于逻辑回归研究了2022届衡阳市高职毕业生对公司的满意度、月均考核评分、月均项目数量、月均工作时长4个工作特征与毕业一年内是否发生离职行为的相关关系。研究结果表明，2022届衡阳市高职毕业生毕业一年内发生过离职行为的比例为53.49%，毕业生对公司的满意度、月均考核评分与发生离职行为存在负相关关系，毕业生月均项目数量、月均工作时长与发生离职行为存在正相关关系。

关键词

逻辑回归，高职毕业生，离职

A Study on the Resignation of Vocational College Graduates in Hengyang City Based on Logistic Regression

Lingyun Li, Jiaodi Shen, Hejun Li, Jiaoli Peng, Xing Sun, Youlin Luo, Heng Chen

Ecological Livable College, Hunan Polytechnic of Environment and Biology, Hengyang Hunan

Received: Mar. 17th, 2024; accepted: Apr. 11th, 2024; published: Apr. 18th, 2024

Abstract

In recent years, over 40% of fresh vocational college graduates in China have resigned within six months each year, which should be widely concerned by all parties. This article distributed a sur-

文章引用: 李凌云, 申娇娣, 李合军, 彭姣丽, 孙兴, 骆又麟, 陈恒. 基于逻辑回归的衡阳市高职毕业生离职问题研究[J]. 应用数学进展, 2024, 13(4): 1338-1344. DOI: 10.12677/aam.2024.134124

vey questionnaire to the 2022 vocational college graduates in Hengyang City, and based on logistic regression, studied the correlation between four job characteristics of the 2022 vocational college graduates: satisfaction with the company, monthly average assessment score, monthly average number of projects, and monthly average working hours, and whether they have resigned within one year of graduation. The research results show that 53.49% of vocational college graduates in Hengyang City in 2022 have experienced resignation behavior within one year after graduation. There is a negative correlation between the satisfaction of graduates with the company, monthly assessment scores, and the occurrence of resignation behavior. There is a positive correlation between the number of monthly projects and average working hours of graduates and the occurrence of resignation behavior.

Keywords

Logistic Regression, Vocational College Graduates, Resignation

1. 引言

党的二十大明确指出，必须坚持“人才是第一资源”，深入实施“人才强国战略”，坚持“人才引领驱动”，强化“就业优先政策”，健全“就业促进机制”，促进“高质量充分就业”，高校毕业生是国家宝贵的人才资源，是促进就业的重要群体[1]。

截至 2024 年 3 月，2022 届衡阳市高职毕业生已工作了一年多的时间，本文通过逻辑回归分析公司满意度、员工考核评分、项目数量和工作时长等数值型工作特征对于 2022 届衡阳市高职毕业生工作一年内离职的影响，为企业的员工管理策略、学校的就业工作和毕业生本人的就业决策提供理论依据。

2. 衡阳市高职院校 2022 届毕业生就业数据

为了解衡阳市高职院校 2022 届毕业生毕业一年内的就业情况，2023 年 9 月至 11 月通过邮箱、班级 QQ 群等途径向 2022 届衡阳市高职毕业生发放问卷星平台调查问卷链接，问卷着重调查对公司的满意度、月均考核评分、月均项目数量和月均工作时长等数值型工作特征以及是否离职。其中，是否离职是一个二分类问题，使用 0 表示未出现离职行为，使用 1 表示出现过离职行为。

最终回收有效问卷 2520 份，出现离职行为的 2022 届衡阳市高职毕业生 1348 人，毕业一年内的离职率为 53.49%。问卷回收后使用 Python 软件进行数据采集、数据清洗和统计分析，计算得到数值型工作特征的最大值、最小值、中位数、均值、标准差、方差、偏度和峰度，计算结果如表 1 所示。

Table 1. Descriptive statistics for numeric features
表 1. 数值型特征的描述统计量

特征	最小值	最大值	均值	标准差	偏度	峰度
对公司的满意度	9	100	61.28	24.86	0.476	−0.670
月均考核评分	36	100	71.61	17.11	0.027	−1.239
月均项目数量	2	7	3.8	1.233	−0.338	−0.495
月均工作时长	96	310	201	49.943	−0.052	−1.134

可通过表 1 中的数值型特征的描述性统计量对 2022 届衡阳市高职毕业生毕业一年内的就业情况有一个初步了解：在对公司的满意度方面，最低的满意度是 9，最高的满意度是 100，均值是 61.28，标准差是 24.86，偏度是 0.476，峰度是−0.670，说明毕业生对公司的满意度集中在较低的区间内；在月均考核评分方面，最低的评分是 36，最高的评分是 100，均值是 71.61，标准差是 17.11，偏度是 0.027，峰度是−1.239，说明毕业生的月均考核评分集中在较低的区间内；在月均项目数量方面，最少的项目数量是 2，最多的项目数量是 7，均值是 3.8，标准差是 1.233，偏度是−0.338，峰度是−0.495，说明毕业生月均项目数量集中在较高的区间内；在月均工作时长方面，最少的工作时长是 96，最多的工作时长是 310，均值是 201，标准差是 49.943，偏度是−0.052，峰度是−1.134，说明毕业生月均工作时长集中在较高的区间内。

3. 基于逻辑回归的特征重要性分析

3.1. 逻辑回归

逻辑回归的思想来源于线性回归[2] [3] [4]，本文研究的问题包含对公司的满意度、月均考核评分、月均项目数量和月均工作时长四个数值型特征，因此可用四元线性回归公式表示为：

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \theta_3 x_3^{(i)} + \theta_4 x_4^{(i)} \tag{1}$$

式中：右上角的(i)表示第 i 份问卷，共有 2520 份有效问卷；$x_1^{(i)}$ 表示第 i 份问卷中的对公司的满意度，$x_2^{(i)}$ 表示第 i 份问卷中的月均考核评分，$x_3^{(i)}$ 表示第 i 份问卷中的月均项目数量，$x_4^{(i)}$ 表示第 i 份问卷中的月均工作时长；θ_1 表示毕业生基于公司满意度做出离职决策的权重参数，θ_2 表示毕业生基于月均考核评分做出离职决策的权重参数，θ_3 表示毕业生基于月均项目数量做出离职决策的权重参数，θ_4 表示毕业生基于月均工作时长做出离职决策的权重参数，其值越大表示其对应的特征重要程度越强；θ_0 表示偏置参数；$\hat{y}^{(i)}$ 表示第 i 份问卷数据中的 4 个特征与权重、偏置通过公式(1)计算得到的输出，是线性回归的预测输出数据，其值可能是任一实数。可用 $y^{(i)}$ 表示第 i 份问卷中的是否离职这一字段的数据，即第 i 份问卷的实际输出数据，实际输出数据 $y^{(i)}$ 与预测输出数据 $\hat{y}^{(i)}$ 之间的差值称为第 i 份问卷的预测误差。

令：

$$\Theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}, X^{(i)} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \tag{2}$$

式中：Θ 是参数矩阵，表示所有参数；$X^{(i)}$ 是特征矩阵，表示处理后的第 i 份问卷中的所有特征。则式(1)可利用矩阵乘法简化为：

$$\hat{y}^{(i)} = \Theta^{\mathrm{T}} \bullet X^{(i)} \tag{3}$$

式中：Θ^{T} 是参数矩阵 Θ 的转置，$\bullet$ 表示矩阵乘法。

为了将线性回归的预测输出数据 $\hat{y}^{(i)}$ 与发生离职行为的概率联系起来，可使用 *sigmoid* 函数可将任一实数 $\hat{y}^{(i)}$ 映射成(0,1)之间的数，实现数值到概率的转换。将线性回归的预测输出数据 $\hat{y}^{(i)}$ 经过 *sigmoid* 函数变换后可得：

$$sigmoid\left(\hat{y}^{(i)}\right) = sigmoid\left(\Theta^{\mathrm{T}} X^{(i)}\right) = \frac{1}{1+\mathrm{e}^{-\Theta^{\mathrm{T}} X^{(i)}}} \tag{4}$$

式中：使用了 *sigmoid* 函数将实数 $\hat{y}^{(i)}$ 映射成(0, 1)之间的数，同时也将线性回归的预测输出数据 $\hat{y}^{(i)}$ 变成

了逻辑回归的预测数据 $sigmoid\left(\Theta^{\mathrm{T}}X^{(i)}\right)$，即通过第 i 份问卷的 4 个特征数据预测其发生离职行为的概率为 $sigmoid\left(\Theta^{\mathrm{T}}X^{(i)}\right)$，未发生过离职行为的概率为 $1-sigmoid\left(\Theta^{\mathrm{T}}X^{(i)}\right)$，可用公式表示为：

$$p^{(i)}=\left(sigmoid\left(\Theta^{\mathrm{T}}X^{(i)}\right)\right)^{y^{(i)}}\left(1-sigmoid\left(\Theta^{\mathrm{T}}X^{(i)}\right)\right)^{1-y^{(i)}} \tag{5}$$

式中：当第 i 份问卷是否离职字段的取值为代表发生离职行为的 1 时，此时第 i 份问卷的实际输出数据 $y^{(i)}$ 取值为 1，则发生离职行为的概率 $p^{(i)}=\left(sigmoid\left(\Theta^{\mathrm{T}}X^{(i)}\right)\right)$；当第 i 份问卷是否离职字段的取值为代表没有发生离职行为的 0 时，此时第 i 份问卷的实际输出数据 $y^{(i)}$ 取值为 0，则发生离职行为的概率 $p^{(i)}=1-sigmoid\left(\Theta^{\mathrm{T}}X^{(i)}\right)$。

为找到某个参数矩阵 Θ 使得全体就业问卷数据的预测误差最小，可采用基于逻辑回归的梯度下降方法求解参数矩阵 Θ，公式为：

$$\Theta=\Theta-\eta\nabla J(\Theta)=\begin{bmatrix}\theta_0\\ \theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{bmatrix}-\eta\begin{bmatrix}\frac{\partial J(\Theta)}{\partial\theta_0}\\ \frac{\partial J(\Theta)}{\partial\theta_1}\\ \frac{\partial J(\Theta)}{\partial\theta_2}\\ \frac{\partial J(\Theta)}{\partial\theta_3}\\ \frac{\partial J(\Theta)}{\partial\theta_4}\end{bmatrix}=\begin{bmatrix}\theta_0+\frac{\eta}{m}\sum_{i=1}^{m}\left(y^{(i)}-p^{(i)}\right)\\ \theta_1+\frac{\eta}{m}\sum_{i=1}^{m}\left(y^{(i)}-p^{(i)}\right)x_1^{(i)}\\ \theta_2+\frac{\eta}{m}\sum_{i=1}^{m}\left(y^{(i)}-p^{(i)}\right)x_2^{(i)}\\ \theta_3+\frac{\eta}{m}\sum_{i=1}^{m}\left(y^{(i)}-p^{(i)}\right)x_3^{(i)}\\ \theta_4+\frac{\eta}{m}\sum_{i=1}^{m}\left(y^{(i)}-p^{(i)}\right)x_4^{(i)}\end{bmatrix} \tag{6}$$

式中：m 取值为 2520，表示有效调查问卷的数量；$\partial J(\Theta)$ 表示逻辑回归的损失函数；$\partial J(\Theta)/\partial\theta_0$、$\partial J(\Theta)/\partial\theta_1$、$\partial J(\Theta)/\partial\theta_2$、$\partial J(\Theta)/\partial\theta_3$、$\partial J(\Theta)/\partial\theta_4$ 分别表示损失函数 $J(\Theta)$ 在不同偏置方向上的偏导数，它们汇聚而成的矩阵就是损失函数的梯度 $\nabla J(\Theta)$；η 被称为学习率，是梯度下降算法中的参数。重复使用式(6)更新参数矩阵 Θ，可以得到预测误差最小的参数矩阵 Θ。

3.2. 评估指标

逻辑回归的常用评估指标有准确率、精确率、召回率、ROC 曲线和 AUC 值，且这些评估指标的计算都依赖于混淆矩阵[5] [6] [7] [8]，本文研究的基于逻辑回归分析离职问题的混淆矩阵如表 2 所示。

Table 2. Confusion matrix based on logistic regression analysis of resignation issues
表 2. 基于逻辑回归分析离职问题的混淆矩阵

真实情况	逻辑回归预测结果	
	出现离职行为	未出现过离职行为
出现离职行为	*TP*	*FN*
未出现过离职行为	*FP*	*TN*

其中，*TP* 表示逻辑回归预测结果与真实情况均未出现离职行为的样本数量；*FP* 表示逻辑回归预测结果为出现离职行为而真实情况为未出现过离职行为的样本数量；*FN* 表示逻辑回归预测结果为未出现过离职行为而真实情况为出现离职行为的人数；*TN* 表示逻辑回归预测结果与真实情况均为未出现过离职行为的人数。

逻辑回归准确率的计算公式为：

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{7}$$

式中：$accuracy$ 表示逻辑回归准确率，$TP + TN$ 表示预测正确的样本数量；$TP + FN + TN + FP$ 表示本研究样本的总数量，即 2520 份样本；准确率 $accuracy$ 表示逻辑回归预测正确的样本数量占总样本的比例。

逻辑回归精确率的计算公式为：

$$precision = \frac{TP}{TP + FP} \tag{8}$$

式中：TP 表示逻辑回归预测结果与真实情况均未出现离职行为的样本数量；$TP + FP$ 表示 2520 份样本中逻辑回归预测结果为出现离职行为的样本数量；精确率 $precision$ 表示逻辑回归正确预测离职的样本数量占逻辑回归预测结果为出现离职行为的样本数量的比例。

逻辑回归召回率的计算公式为：

$$recall = \frac{TP}{TP + FN} \tag{9}$$

式中：TP 表示逻辑回归预测结果与真实情况均未出现离职行为的样本数量；$TP + FN$ 表示真实情况为出现离职行为的样本数量；召回率 $recall$ 表示逻辑回归正确预测离职的样本数量占实际出现离职行为的样本数量的比例。

逻辑回归的 ROC 曲线绘制依赖于不同阈值下的真正例率和假正例率，真正例率和假正例率分别定义为：

$$\begin{cases} TRP = \dfrac{TP}{TP + FN} \\ FRP = \dfrac{FP}{TN + FP} \end{cases} \tag{10}$$

式中：真正例率 TRP 与式(9)中的召回率 $recall$ 相同，表示逻辑回归正确预测为离职的样本数量占实际出现离职行为的样本数量的比例；FP 表示逻辑回归预测结果为出现离职行为而真实情况为未出现过离职行为的样本数量；$TN + FP$ 表示真实情况为未出现离职行为的样本数量；假正例率 FRP 表示逻辑回归错误预测为离职的样本数量占实际未出现离职行为的样本数量的比例。

逻辑回归的 AUC 值被定义为 ROC 曲线下的面积，当 AUC 值在 0.8 以上时就表示逻辑回归预测效果很好[9] [10]。

3.3. 基于逻辑回归的特征重要性分析实验步骤

基于逻辑回归的衡阳市高职毕业生离职特征重要性实验步骤如下：

1) 使用符合高斯分布的随机数初始化参数矩阵 Θ；

2) 对数值型数据进行标准化处理；

3) 计算得到预测误差最小的参数矩阵 Θ，如公式(6)所示；

4) 根据逻辑回归预测结果计算混淆矩阵，如表 2 所示；

5) 计算得到逻辑回归的准确率、精确率和召回率，如公式(7)、公式(8)和公式(9)所示；

6) 计算不同阈值下的真正例率和假正例率，并将它们分别作为纵轴和横纵绘制 ROC 曲线，如公式(10)所示；

7) 计算 ROC 曲线下的面积，即 AUC 值；

8) 结合参数矩阵 Θ 和评估指标分析衡阳市高职毕业生离职特征重要性。

4. 实验

根据实验步骤，基于逻辑回归的特征重要性分析实验图像与实验数据分别如图 1 和表 3 所示。

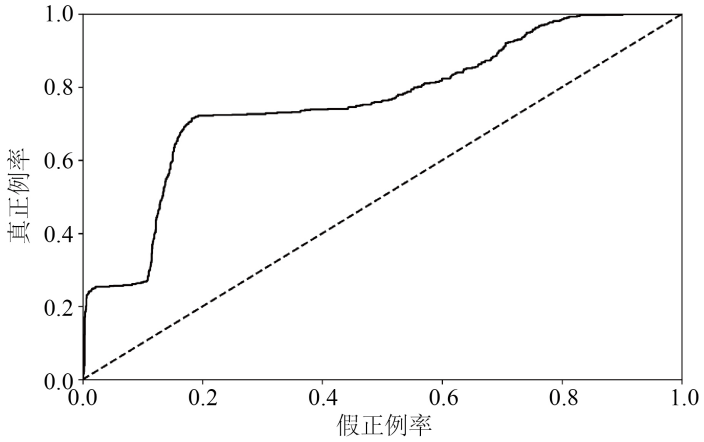

Figure 1. ROC curve of resignation issues for vocational college graduates in Hengyang city based on logistic regression
图 1. 基于逻辑回归的衡阳市高职毕业生离职问题 ROC 曲线

图中：ROC 曲线靠近左上角说明使用逻辑回归对于本文研究的衡阳市高职毕业生离职问题有较好的预测性能。

Table 3. Experimental data of feature importance analysis based on logistic regression
表 3. 基于逻辑回归的特征重要性分析实验数据

参数矩阵 Θ					准确率	精确率	召回率	AUC 值
θ_0	θ_1	θ_2	θ_3	θ_4				
−1.40	−1.04	−0.15	0.31	0.23	0.80	0.73	0.77	0.76

表中：偏置参数 θ_0 是−1.40；权重参数 θ_1 和 θ_2 分别是−1.04、−0.15，表示毕业生对公司的满意度、月均考核评分与发生离职行为存在负相关关系，且对公司的满意度这一特征的负相关程度较强；权重参数 θ_3 和 θ_4 分别是 0.31、0.23，表示毕业生月均项目数量、月均工作时长与发生离职行为存在正相关关系，且月均项目数量这一特征的正相关程度较强；基于逻辑回归的衡阳市高职毕业生离职问题的准确率是 0.80、精确率是 0.73、召回率是 0.77、AUC 值是 0.76，这些评估指标均说明使用逻辑回归对于本文研究的衡阳市高职毕业生离职问题有较好的预测性能，由此得到的各个特征的重要性也有着很高的可信度。

5. 结论与启示

本文以 2520 份有效问卷数据为依据，基于逻辑回归研究了 2022 届衡阳市高职毕业生对公司的满意度、月均考核评分、月均项目数量、月均工作时长 4 个工作特征与毕业一年内是否发生离职行为的相关关系，结论与启示如下：

1) 毕业生对公司的满意度和月均考核评分越高，毕业生做出离职行为的可能性就越低，且对公司的

满意度的重要性更大；毕业生月均项目数量和月均工作时长越高，毕业生做出离职行为的可能性也越高，且月均项目数量的重要性更大；评估指标数据表明得到的各个特征的重要性有较高的可信度。

2) 避免毕业生离职需要企业在毕业生本人的主观情感方面做出努力，包括确立企业文化和价值观、提供良好的工作环境和氛围、公平合理的薪酬和福利、提供良好的职业发展空间。这些策略的实施有利于提高毕业生对公司的满意度，毕业生对公司的满意度。

基金项目

2023 年衡阳市社会科学基金项目“基于逻辑回归的衡阳市高职毕业生离职问题研究”(2023D023)。

参考文献

[1] 习近平. 高举中国特色社会主义伟大旗帜为全面建设社会主义现代化国家而团结奋斗——在中国共产党第二十次全国代表大会上的报告[J]. 中华人民共和国国务院公报, 2022(30): 4-27.

[2] 马庆祥, 杨娟美. 高职毕业生“高就业率”“高离职率”现象研究[J]. 安徽电气工程职业技术学院学报, 2020, 25(2): 89-93.

[3] 霍燕. 烟草行业新入职高校毕业生离职原因及对策分析[J]. 经济师, 2023(2): 282+284.

[4] 陈姣姣, 肖雪, 范腾阳, 等. 贵州省 304 名订单定向医学毕业生工作价值观对离职意愿影响的调查[J]. 预防医学情报杂志, 2022, 38(11): 1460-1465.

[5] 徐平利. 职业院校毕业生“就离职”现象研究[J]. 江苏高职教育, 2022, 22(2): 56-62.

[6] 高法文. 高职院校毕业生职业适应初期离职情况研究——以深圳职业技术学院为例[J]. 深圳职业技术学院学报, 2019, 18(4): 61-66.

[7] 边玉宁, 陆利坤, 李业丽, 曾庆涛, 孙彦雄. 基于逻辑回归的金融风投评分卡模型实现[J]. 计算机科学, 2020, 47(S2): 116-118.

[8] 方然可, 刘艳辉, 苏永超, 黄志全. 基于逻辑回归的四川青川县区域滑坡灾害预警模型[J]. 水文地质工程地质, 2021, 48(1): 181-187.

[9] 费云利. 计算机逻辑回归分析[J]. 湖南工业职业技术学院学报, 2020, 20(1): 14-17.

[10] 邓飞, 张荣稳, 余靖, 等. 基于 ROC 曲线的逻辑回归切割值寻优方法研究及应用[J]. 昆明理工大学学报(自然科学版), 2023, 48(1): 53-57.