文章编号：2095-2163(2023)06-0162-06 中图分类号：F272.92;TP181 文献标志码：A

基于 Logistic 回归与决策树的员工数据可视化与离职预测研究

龚建伟，张林锋，佘奇根，于 放
（北京师范大学 香港浸会大学联合国际学院理工科技学院，广东 珠海 519087）

摘 要：通过对员工离职的数据集进行可视化，可以发现员工的离职与许多因素有关，这意味着运用机器学习方法对员工离职进行预测是可行的。实验对数据集中的文本内容进行赋值，并计算了不同因素与离职与否之间的相关系数。基于赋值后的数据集，构建了逻辑回归模型，该模型可以根据给定的员工情况输出员工离职的概率。经对 7 种模型的优化与对比结果表明，优化后的决策树算法的准确性最高。

关键词：Logistic 回归；决策树；机器学习；员工离职预测

Research on employee data visualization and turnover prediction based on logistic regression and decision tree

GONG Jianwei, ZHANG Linfeng, SHE Qigen, YU Fang

(Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai Guangdong 519087, China)

【Abstract】 By visualizing the data set of employee turnover, we can find that employee turnover is related to many factors, which means that it is feasible to use machine learning method to predict employee turnover. Based on collated datasets, a logistic regression model is constructed, which can output the probability of employee turnover according to the given employee situation. By comparing with several machine learning models, we found that the optimized decision tree has the highest accuracy.

【Key words】 logistic regression; decision tree; machine learning; employee turnover prediction

0 引 言

对于企业来说，员工离职率高意味着企业难以留住人才，同时也给未来的经营带来许多不确定性。究其员工为什么会离职，以及不同员工的离职概率等，都是企业不得不面临的难题。搞清楚与员工离职有关的因素不仅可以帮助企业预测未来的人力资源变动情况与需求，同时有助于帮助企业找到员工离职背后的原因。显然，员工的离职并不完全是随机的，员工的自身情况与工作条件等诸多原因都可能对其离职概率产生影响。因此，使用机器学习方法对员工离职的概率进行研究具有充分的可行性，企业也可运用这些方法来建立员工离职预警机制[1-2]，这对企业的长远发展可谓裨益良多。

早期研究人员依据历史数据进行预测，往往是将传统统计方法用于时间序列模型当中，如 ARIMA[3]、多元线性回归模型等。后来，研究人员逐渐开始运用机器学习算法来对股票等信息进行预测，这些算法相较传统模型而言效果通常来说要更令人满意[4]。离职预测问题是一个典型的分类问题，目前已有诸多机器学习算法可以应用于此类问题[5]，完全可以应用于离职预测当中。虽然有许多算法可以利用，但是不同算法在准确度和样本规模适应性等方面各有千秋[6-7]。目前，已有一些研究者利用 XGBoost 算法[8]、随机森林[9-10]等方法进行了员工离职预测模型的构建，但这些研究大多仅基于一种算法，亦无法给出离职的概率而只能进行简单的是非判断。也有一些研究者对不同算法的准确性进行了对比[11]，但该研究所对比的模型均使用默认参数而没有进行优化，故模型仍有值得改进之处。在运用机器学习算法进行预测时，需要警惕过拟合的问题，此类问题可以通过参数调优来解决[12]。

作者简介：龚建伟(1998-)，男，硕士，助理工程师，主要研究方向：人工智能与数码媒体。
通讯作者：龚建伟 Email：jianwei_gong_usts@163.com

收稿日期：2022-06-07

1　相关技术

1.1　数据可视化

在实验正式开始之前，需要运用 dropna 函数来清洗实验数据，该函数可以去除数据集中含有缺失值的数据行，进而确保在后续可视化与用于预测的数据都是有效的。可视化部分主要借助 matplotlib 绘图库来展示数据集的基本信息，该绘图库可以用于绘制饼状图、条形图等图片。完成基本的可视化之后，为了进一步判断不同因素与离职之间的相关性大小，实验运用了 DataFrame 内建的 corr 函数，该函数可以用于计算不同数据之间的相关系数。为了更加直观地看出不同因素与离职之间相关系数的正负与大小如何，实验同样以可视化的形式展示了相关系数的条形图。

1.2　离职预测

在预测部分，首先以 Logistic 回归方法对员工离职与否进行了预测，该方法的特色在于可以给出员工离职的概率。通过测试可以得到 Logistic 回归的准确率和 *ROC* 曲线、*AUC* 值等指标，这些指标可以用于判断模型的优良程度。凭借搭建好的 Logistic 回归模型，可以构建根据员工个人信息来预测离职概率的模块。随后，实验使用 sklearn 库中的多种分类器对员工离职与否进行了预测，测试了 K 近邻算法、决策树、随机森林、极度决策树、梯度提升分类器、AdaBoostClassifier 和支持向量分类器等多种模型的准确率，最终采用其中预测准确度最高的模型，以构建判断员工是否会离职的预测模块。

2　数据可视化

2.1　实验环境

实验所选硬件环境为 16GB 计算机内存，Windows10 64 位 1909 版操作系统，搭载有 Intel Xeon E3-1231v3 3.4 GHz 处理器与 GTX960 显卡；软件环境为基于 Python3.6 编程语言的 sklearn 机器学习库中集成的 linear_model 与 K 近邻算法、决策树等模型。

2.2　数据集

2.2.1　数据集导入

实验数据来源于 kaggle 上一份某印度公司的人力资源部门提供的约有四千余条数据的数据集，其主要内容包括员工的个人信息有：教育程度、入职年份、所在城市、收入水平、年龄、性别、是否被冷落过、工作经验以及未来两年内离职与否等。

将数据集导入之后，通过 head 函数可以观察到数据集所包含的内容，如图 1 所示。

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	Bachelors	2017	Bangalore	3	34	Male	No	0	0
1	Bachelors	2013	Pune	1	28	Female	No	3	1
2	Bachelors	2014	New Delhi	3	38	Female	No	2	0
3	Masters	2016	Bangalore	3	27	Male	No	5	1
4	Masters	2017	Pune	3	24	Male	Yes	2	1
5	Bachelors	2016	Bangalore	3	22	Male	No	0	0
6	Bachelors	2015	New Delhi	3	38	Male	No	0	0
7	Bachelors	2016	Bangalore	3	34	Female	No	2	1
8	Bachelors	2016	Pune	3	23	Male	No	1	0
9	Masters	2017	New Delhi	2	37	Male	No	2	0

图 1　数据集基本情况

Fig. 1　Dataset overview

2.2.2　数据可视化分析

为了更加直观地了解员工的大体状况，调用 matplotlib.pyplot 对员工的基本信息进行了可视化。图 2 为员工入职时间的可视化展示，其它针对工作经验、学历等方面的展示大体相同，故不再赘述。

基本的可视化完成之后，通过 value_counts 函数计量了根据不同要素划分的员工群体离职比例，以观察哪些群体的员工更有离职的可能。图 3 中的百分比均通过 round 函数保留小数点后两位，每个群体标题之后所跟的百分比为该群体在整体员工中所占比例。

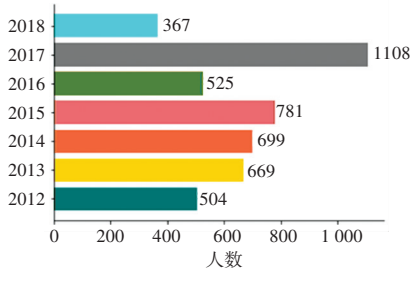
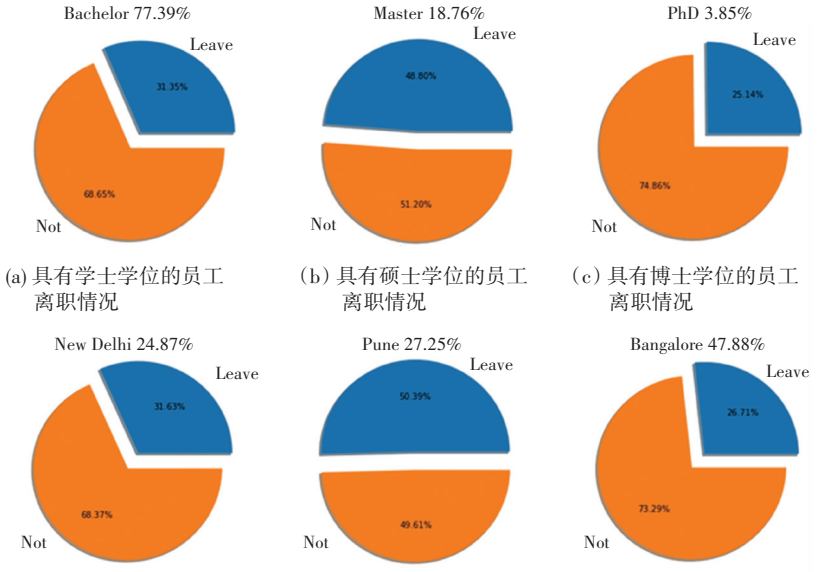


图 2　员工入职年份的可视化

Fig. 2　Visualization of the joining year of employees



(a) 具有学士学位的员工离职情况 (b) 具有硕士学位的员工离职情况 (c) 具有博士学位的员工离职情况

(d) 新德里地区员工离职情况 (e) 浦那地区员工离职情况 (f) 班加罗尔地区员工离职情况

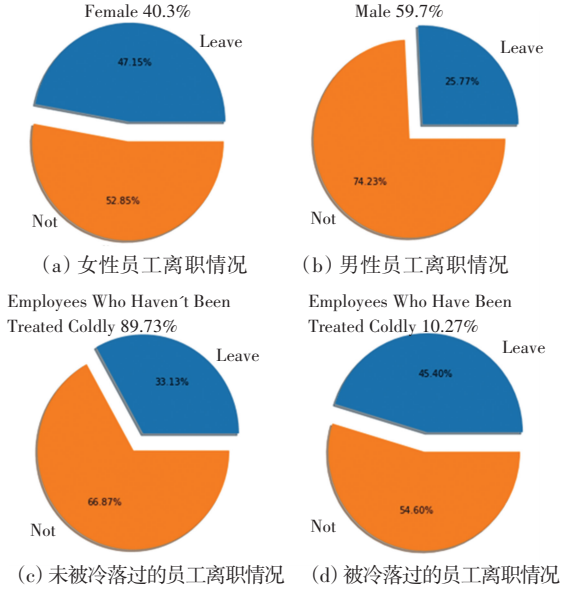
图 3 根据学历或者地区划分的不同群体员工离职比例

Fig. 3 Turnover ratio of employees in different groups by education background or region

通过数据可视化可以看出,员工的离职并非完全随机,不同群体的员工离职率确有显著差异。通常情况下,人们会认为学历越高的员工工作会更加稳定,因为这意味着他们的专业技术水平更高,但数据集展现出的情况并非如此。究其原因,可能是硕士学历持有者在求职市场中属于相对较为稀少的高级人才,同时也是企业高级人才构成的中坚力量,在就业市场非常受欢迎。不仅如此,硕士学历持有者的年龄往往也更年轻,相比博士学历者更有优势,因此其可能会选择跳槽来换取更高的待遇。本科学历的员工跳槽不一定能有更好待遇,所以离职的会更少;博士学历者属于稀缺人才(占比不到 4%),公司对其待遇和许诺的前途都会更好,所以离职率也会偏低。当然,年龄同样也是一个影响因素,博士毕业者大多在三十岁左右,其年龄上没有什么优势可言,跳槽的风险也可能更大,这对离职率亦有影响。

在地区方面,实验选取的数据集中的员工分布于 3 个城市,其中新德里(印度首都)是2 500万人口规模的城市,班加罗尔(印度第三大城市)约1 000万人口规模的城市,浦那(印度西部城市)则是约 500 万人口规模的城市。可以发现,人口规模最小的城市离职率最高,新德里的员工离职率相对偏高,而人口规模处于中位城市的员工离职率最低。究其原因,城市的人口规模小可能意味着该地薪资水平不高、工作环境不好、未来发展受限等;但大城市往往也伴随着高生活成本、日常通勤时间长、竞争激烈等问题,这都会带来一定的压力,所以人口规模适中的城市离职率反而最低。

地区和学历已被证实对离职概率有显著影响,故实验进一步绘制了根据性别和是否被冷落过划分的员工群体离职状况,如图 4 所示。



(a) 女性员工离职情况 (b) 男性员工离职情况

(c) 未被冷落过的员工离职情况 (d) 被冷落过的员工离职情况

图 4 不同群体员工离职比例

Fig. 4 Turnover ratio of employees in different groups

2.2.3 相关系数可视化

上述数据可视化结果仅能证明了员工自身与周围的诸多因素对离职与否有影响,并不能证明影响的程度如何,不同因素对离职的影响大小仍需以相关系数的形式展现。为了计算相关系数,实验对数据集中的文本信息进行了赋值,即对不同学历、不同

地区分别按照一定顺序(如:学历从低到高)赋值 1、2、3 等,随后计算了相关系数矩阵,并用 matplotlib.pyplot 绘制了其他因素与离职与否的相关系数,如图 5 所示。

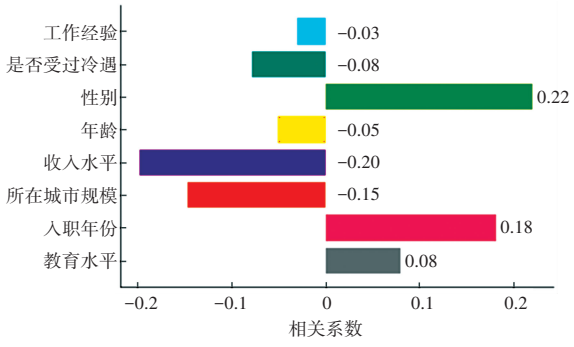


图 5　不同因素与离职与否的相关系数

Fig. 5　Correlation coefficient between different factors and turnover

3　离职预测

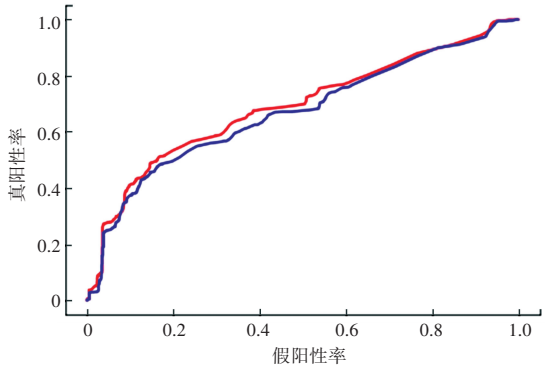
3.1　基于 Logistic 回归的离职概率预测

Logistic 回归是一种经典的预测方法,其原理是将线性回归的结果带入 Sigmoid 函数,从而使连续变量转换为 0~1 区间的一个概率值。当概率大于 0.5 时,样本归为正,当概率小于 0.5 时,样本归为负[13]。这一特性意味着逻辑回归模型可以被用于概率预测。Sigmoid 函数如式(1)所示:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

实验调用了 sklearn.linear_model 来进行 Logistic 回归,并将数据集的前3 000条数据作为训练集,剩余数据作为测试集。为了便于训练,在训练之前需要将数据集中的文本信息转换为数字。由于年龄与加入年份的数据与其他类型数据有较大偏差,故在 Logistic 回归时仅选择教育程度、收入水平、城市、性别、工作经验以及是否被冷落过 6 项来进行预测。对于 Logistic 回归模型来说,*ROC* 曲线与 *AUC* 值是相当重要的指标,如果得出的 *AUC* 值小于等于 0.5,则说明预测并不可行。因此,实验用 sklearn.metrics 绘制了 *ROC* 曲线,如图 6 所示。

Logistic 回归模型的 *AUC* 值越接近 1 则说明模型越优秀[14],0.67 左右这个数值只能说是差强人意,但仍有利用价值。通过.predict()方法与进一步计算实验发现,Logistic 回归模型的准确率为 72.96%。准确率不高的原因可能与数据集本身的内容有一定关联。此外,将入职年份排除在模型之外,也可能对 Logistic 回归的准确率造成了相当大的影响,因为从相关系数来看,入职年份与离职与否之间存在相当大的相关性。在 Logistic 回归模型训练完成之后,仅需将数值替换为需要输入的内容即可构建员工离职概率预测模块,如图 7 所示。



训练集 *AUC* 值:0.690 218 845 660 895 8

测试集 *AUC* 值: 0.669 453 288 213 425 2

图 6　Logistic 回归模型的 *ROC* 曲线与 *AUC* 值

Fig. 6　*ROC* curve and *AUC* value of logistic regression model

```
请输入教育程度(1:学士;2:硕士;3:博士): 2
请输入所在城市: 1:浦那;2:班加罗尔;3:新德里2
请输入薪资水平: 1:低;2:中;3:高3
请输入性别: 1:男;2:女1
请输入是否坐过冷板凳: 1:有2:无2
请输入工作经验年限: 0-73
这位员工在未来两年内离职的概率为: 0.296 432 357 036 412 1
```

图 7　基于 Logistic 回归的员工离职概率预测模块

Fig. 7　Prediction of turnover probability based on logistic regression

在图 7 中,用户输入了一名硕士学历、在班加罗尔工作、高薪资水平、男性、没有被冷落过、三年工作经验的员工信息,基于 Logistic 回归的预测模块给出了该员工离职的概率为 29.64%。由此可见,模型仅需输入员工的相关信息即可给出该员工的离职概率,若有更多用于训练的数据,则准确度也可进一步提高,说明该模型具有较高的实用价值。在现实生活当中,员工的信息可能更为多元,但预测的原理是一致的,故预测模型的可迁移性亦有保证。

3.2　基于机器学习算法的员工离职预测

3.2.1　模型对比

Sklearn 库中有许多模型可供使用,这些模型大多属于分类器,无法给出离职概率,只能针对员工离职与否进行分类判断。在此测试中,将数据集中 90%划分为训练集,10%划分为测试集,random_state 设置为 39。实验共对 7 种模型进行了测试,下面是对 KNN 模型的测试。

在正式测试 KNN 模型之前,需要针对不同 *K* 值(即临近邻居的数量)进行测试以选择最优化的结果。图 8 中的测试结果表明,*K* = 11 时的 KNN 模型

具有最佳的准确率(81.55%)。这一准确率高于 Logistic 回归,同时也为调优后的结果,在对其他模型进行测试时,也会采取类似的调优操作。

以决策树模型为例,在未调优时,测试结果如图 9 所示,准确度约为 85.2%。

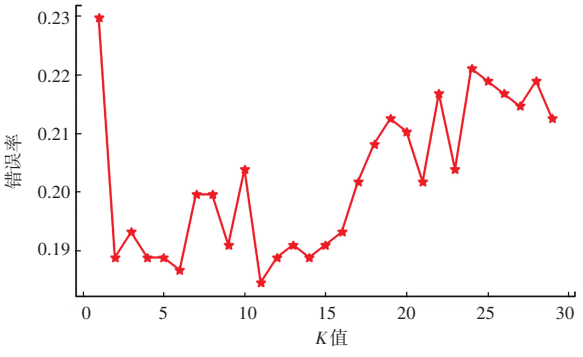


图 8 不同 *K* 值的 KNN 模型错误率

Fig. 8 KNN model error rate with different *K* values

```
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier( random_state=42, criterion='entropy', splitter='random')
tree.fit(x_train, y_train)
y_predict = tree.predict(x_test)
print("决策树的准确率为: ",tree.score(x_test,y_test))
```

```
决策树的准确率为: 0.851 931 330 472 103
```

图 9 未调优时决策树的测试结果

Fig. 9 Test results of decision tree without tuning

对于决策树模型来说, max_depth(树的最大深度)、min_samples_leaf(叶节点必须有的最小样本数量)和 min_samples_split(前节点允许分裂的最小样本数)3 个参数的设置会对准确度产生明显的影响,如果设置不当的话,准确度反而会下降。如:将参数设置为 max_depth = 8、min_samples_leaf = 2、min_samples_split = 7 时,测试结果如图 10 所示,准确度约为 84.8%,这一结果甚至要劣于未调优的决策树模型。

```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier(criterion='entropy',max_depth=8, min_samples_leaf=2, min_samples_split=7)
model.fit(x_train, y_train)
pred = model.predict(x_test)
scores.append({
    'model' : 'DecisionTreeClassifier',
    'score' :  model.score(x_test,y_test),
'f1_score'  :  f1_score(y_test , pred)
})
print("max_depth=8, min_samples_leaf=2, min_samples_split=7时,决策树的准确率为: ",model.score(x_test,y_test))
```

```
max_depth=8, min_samples_leaf=2, min_samples_split=7时,决策树的准确率为: 0.8476394849785408
```

图 10 调优失败时决策树的测试结果

Fig. 10 Test results of decision tree when tuning fails

为了找到合适的调优参数,实验利用网格搜索法(GridSearchCV)来寻找最优的调优参数,该方法分为网格搜索和交叉验证两部分,能够在验证集上找到准确度最高的参数。最终的调优参数寻找结果如图 11 所示。

```
from sklearn.model_selection import GridSearchCV
grid_params = {
    'criterion' : ['gini', 'entropy'],
    'max_depth' : [3, 5, 7, 10],
    'min_samples_split' : range(2, 12, 1),
    'min_samples_leaf' : range(2, 12, 1)
}

grid_search = GridSearchCV(tree, grid_params, cv = 5, n_jobs = -1, verbose = 1)
grid_search.fit(x_train, y_train)
tree = grid_search.best_estimator_
y_pred = tree.predict(x_test)
print("调优参数是:",grid_search.best_params_)
```

```
Fitting 5 folds for each of 800 candidates, totalling 4000 fits
调优参数是: {'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 7}
```

图 11 利用网格搜索法寻找调优参数

Fig. 11 Use GridSearchCV to find tuning parameters

将参数 max_depth = 10、min_samples_leaf = 3、min_samples_split = 7 输入模型后,测试可得该决策树模型的准确度约为 87.3%。在相同的实验环境下,进一步测试了其他几种模型,测试结果如图 12 所示。通过对比发现,调优后的决策树模型有着最高的准确度,因此实验最终选择了该模型。

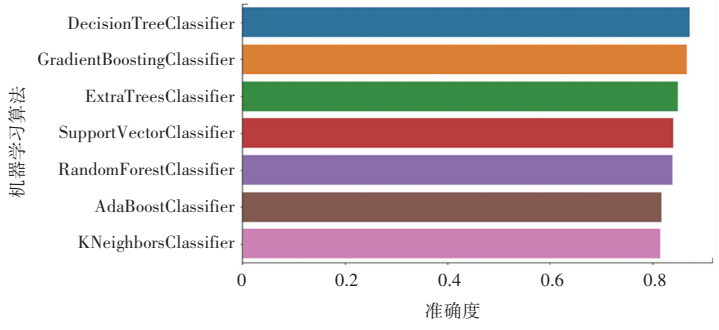


图 12 七种分类器的测试结果

Fig. 12 Test results of seven classifiers

3.2.2 预测模块

由于设置调优参数后的决策树模型具有最高的准确率,因此实验基于该决策树模型搭建了能够根据员工个人信息来判断员工是否会离职的模块,如图 13 所示。

当用户输入了一名硕士学历、2017 年入职、在浦那工作、中等薪资水平、25 岁、女性、被冷落过、1 年工作年限的员工,预测模块随即给出了预测结果:该员工会离职。由此可见,该模型在测试集上有着较高的准确度,因此具有一定的应用价值。

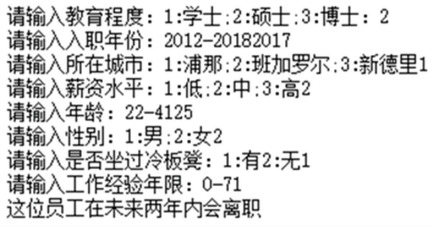


图 13　基于决策树的员工离职与否预测

Fig. 13　Prediction of employee turnover with decision tree

4　结束语

为了搭建员工离职预测模型，实验首先将数据集可视化，以探究与离职有关的种种影响因素；然后运用 Logistic 回归与优化后的决策树模型搭建了员工离职预测模块，分别给出了离职的概率与是否离职的二分类预测。在多种机器学习算法的对比当中，实验对这些模型进行了调优，这意味着对比更加科学，且最终准确率也更高。本实验的美中不足在于 Logistic 回归模型的准确率相对不高，未来研究可考虑对模型进一步改进，将入职年份等与离职有较强相关性的因素纳入模型当中，这对进一步提升模型的准确度会有所帮助。

参考文献

[1] 李芸，胡可，董欣雨，等. 基于 SVM 算法的企业员工离职预警研究[J]. 中国商论，2020，29(6)：20-22.

[2] 吕佳昕. 企业员工离职预测系统的设计与实现[D]. 哈尔滨：东北林业大学，2022.

[3] CHEN Y，HAO Y. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction [J]. Expert Systems with Applications，2017，80：340-355.

[4] 李婧琦. 基于鲸鱼算法优化 LSTM 的股票价格预测模型[J]. 智能计算机与应用，2023，13(2)：35-40.

[5] 杨剑锋，乔佩蕊，李永梅，等. 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.

[6] 万毅斌. 非均衡数据下基于 SMOTE-SVM 的员工离职预测研究[D]. 上海：东华大学，2022.

[7] 吴学亮，娄莉. 样本均衡与特征选择在员工离职倾向预测上的应用[J]. 智能计算机与应用，2022，12(7)：181-184.

[8] 王瑞，尹红，强冰冰. 基于改进 XGBoost 的企业员工离职预测模型[J]. 信息技术，2021，45(8)：12-15，20.

[9] 徐昆，赵东亮. 餐饮连锁店员工离职倾向预测研究[J]. 合作经济与科技，2018，34(8)：170-173.

[10] 陈沛光. 基于随机森林模型的电力企业员工离职倾向预测研究[J]. 化工管理，2018，33(36)：19-20.

[11] 乔源，陈梦帆. 基于多种机器学习算法的员工离职预测模型对比及解释研究[J]. 商讯，2021，39(27)：189-191.

[12] 彭义春，张捷，覃左仕. 基于随机森林算法的职位薪资预测[J]. 智能计算机与应用，2021，11(10)：67-72.

[13] 邹晓辉. 基于 Logistic 回归的数据分类问题研究[J]. 智能计算机与应用，2016，6(6)：139-140，143.

[14] NAKAS C T，YIANNOUTSOS C T. Ordered multiple-class ROC analysis with continuous measurements[J]. Statistics in medicine，2004，23(22)：3437-3449.

(上接第 161 页)

[7] ELBOUKHARI M，AZIZI M，AZIZI A. Analysis of quantum cryptography protocols by model checking[J]. Int. J. Universal Comput. Sci，2010，1：34-40.

[8] 路松峰，陈莹，胥永康，等. 基于 QCircuit 的 BB84 窃听仿真与分析[J]. 计算机学报，2011，34(2)：229-235.

[9] 朱丽娟. 量子密钥分配协议仿真平台的研究与设计[J]. 计算机与数字工程，2012，40(11)：112-114.

[10] 付益兵，孙立炜. BB84 协议及其 MATLAB 仿真[J]. 科技资讯，2014，12(29)：17.

[11] 陈实，吕洪君，解光军.一种 B92 协议量子电路设计与仿真[J]. 量子电子学报，2016，33(1)：51-55.

[12] 孙茂珠. 基于 BB84 协议的量子密钥分发模拟实验[J]. 电子技术与软件工程，2017(1)：90-91.

[13] 周争艳. 量子通信协议的安全性与仿真研究[D]. 北京邮电大学，2020.

[14] ADU-KYERE A，NIGUSSIE E，ISOAHO J. Quantum Key Distribution：Modeling and Simulation through BB84 Protocol Using Python3[J]. Sensors，2022，22(16)：6284.

[15] 刘红，闫凤利. 未知单量子态的传输[J]. 河北师范大学学报(自然科学版)，2007(3)：321-323，340.

[16] Gilles Brassard，PawelHorodecki，Tal Mor. TelePOVM - A generalized quantum teleportation scheme.[J]. IBM Journal of Research and Development，2004，48(1).

[17] 郭宏，李政宇，鹏翔，等. 量子密码[M]. 北京：国防工业出版社，2016：1-565.

[18] 邓富国. 量子通信理论研究[D]. 北京：清华大学，2004.

[19] 卢奉宇，银振强，王双，等. 50 km 无特征源的测量设备无关量子密钥分发实验[J]. 光学学报，2022，42(3)：238-245.