Modern Management 现代管理, 2022, 12(12), 1617-1629
Published Online December 2022 in Hans. http://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2022.1212209

Hans汉斯

基于机器学习的企业员工流失预警分析

柴天潇，钱雨欣，王　文

天津商业大学，天津

收稿日期：2022年10月21日；录用日期：2022年11月15日；发布日期：2022年12月8日

摘　要

企业培养人才需要大量的成本，为了降低企业成本，降低人才流失率，应当注重企业员工流失分析。员工流失分析是评估公司员工流动率的过程，本文基于Kaggle平台分享的数据集，在数据探索性分析和预处理的基础上，采用多种机器学习算法，构建企业员工流失预警模型，并进行模型的比较评价，目的是找到影响员工流失的主要因素，预测未来的员工离职状况，减少重要价值员工流失。

关键词

机器学习，企业员工流失预警，决策树模型，GBDT模型，XGBoost模型

Early Warning Analysis of Enterprise Employee Turnover Based on Machine Learning

Tianxiao Chai, Yuxin Qian, Wen Wang

Tianjin University of Commerce, Tianjin

Received: Oct. 21st, 2022; accepted: Nov. 15th, 2022; published: Dec. 8th, 2022

Abstract

Enterprises need a lot of costs to cultivate talents. In order to reduce the cost of enterprises and the rate of brain drain, we should pay attention to the analysis of employee turnover. Employee turnover analysis is the process of evaluating the company's employee turnover rate. Based on the data set shared on the Kaggle platform, this paper uses a variety of machine learning algorithms to build an early warning model for enterprise employee turnover, compare and evaluate the models. The purpose is to find the main factors that affect employee turnover, predict future employee turnover, and reduce the loss of employees with important values.

文章引用: 柴天潇, 钱雨欣, 王文. 基于机器学习的企业员工流失预警分析[J]. 现代管理, 2022, 12(12): 1617-1629.
DOI: 10.12677/mm.2022.1212209

Keywords

Machine Learning, Early Warning of Employee Turnover, Decision Tree, GBDT Model, XGBoost Model

1. 引言

人才是企业的核心资源，知识水平和技能较高的员工本身具有更强的流动意愿，如果不对其稳定性进行预测，进而采取针对性措施，必然会因为员工的离职给企业带来人才资源的空缺。员工流失预警的关键就是找到员工流失的原因，然后建立科学、有效的员工流失预警分析模型。通过模型识别内外部环境对员工离职的影响因素，对员工的离职情况进行合理的预测，并及时的采取相应对策，减少重要员工流失情况。

本文将 Kaggle 平台分享的数据作为数据集，在数据探索性分析和预处理的基础上进行决策树模型、GBDT 模型、XGBoost 模型 3 种算法模型构建，对企业员工流失进行预测，有利于企业对员工流失所引发的人事危机进行预防处理。

2. 文献综述

2.1. 企业员工流失预警问题研究

赵敏和谭腾飞发表的《网络水军的成因及其发展——以勒温基本公式“B = f(P·E) = f(LSP)”为视角》提到美国心理学家库尔特·勒温(Kurt Lewin, 1890~1947)认为，人的心理活动是一种心理场或是在生活空间里发生的，一个人的行为(B)取决于个人(P)和他的环境(E)的相互作用，也就是说，行为取决于个体的生活空间[1]。

Mohamed Kamal Abdien (2019)发表的《Impact of communication satisfaction and work-life balance on employee turnover intention》探讨了酒店员工沟通满意度与工作生活平衡的关系。结果表明，沟通氛围、主管沟通、同事沟通、组织整合、个人反馈和媒体质量是企业绩效的重要维度，对企业绩效有显著负向影响。研究还发现，工作灵活性、工作自主性和管理者支持是生活平衡的重要维度，对现代质量管理有显著的负向影响[2]。

张艳菊和孙萌(2020)研究了员工压力在顾客不文明行为和职场不文明行为的共同作用下对员工离职倾向的影响，提出顾客的无端、无礼的行为也会在一定程度上造成员工压力的增大。同时，受到职场上不文明行为的影响，也会加大员工的压力。员工的自我压力、顾客不文明行为和职场不文明行为造成的压力都会进一步扩大压力感，造成离职倾向的产生[3]。

张萌(2021)在《员工工作压力对员工离职倾向的影响模型构建》中提出，不同行业或企业对专业技能人才的要求越来越高，但同时企业由于自身招聘制度缺陷、经营成本增加、薪资水平下降等因素，使得优秀专业技能人才不断流失[4]。

2.2. 机器学习方法在现代管理中的运用

王梦针对人力资源管理中的风险预警问题，在《基于支持向量机的人力资源管理风险预警研究》中

提到引入具有小样本机器学习功能的支持向量机 SVM 进行拟合预测，将影响人力资源管理过程的影响因素。按照岗位风险、招聘风险、培训风险、绩效考核风险、薪酬管理风险、职业发展风险与企业文化激励风险 7 个一级指标，同时将一级指标细化为 20 个二级指标作为支持向量机 SVM 的输入变量，并验证了利用支持向量机 SVM 模型作为人力资源管理风险预警的可行性与适用性[5]。

刘春燕在《基于 XGBoost 的员工流失预测研究》一文中介绍了使用集成学习中基于 Bagging 的最具有代表性的随机森林和基于 Boosting 的最强大的 XGBoost，使用 IBM 公司的真实数据构建预测模型，并通过评估模型，挑选出最适合流失数据的模型，提出预测模型具体的应用方案，以及招聘要求、流失预警、员工个人流失方案，供企业具体实践参考[6]。

杨守斌在《基于机器学习方法的 A 公司软件工程师绩效评价研究》一文中对员工工作绩效评价与机器学习理论进行了阐述，归纳和整理了前人关于员工绩效评价指标的理论，构建了 A 公司软件工程师绩效评价的指标体系，通过机器学习技术中的决策树算法构建了软件工程师绩效评价模型，根据所构建的绩效评价体系及模型对 A 公司的部分软件工程师的绩效进行了试评价[7]。

王玲在《机器学习技术在企业智能财务中的应用研究》一文中基于机器学习技术的运作原理，探讨该技术在智能财务中应用的四大场景，即优化会计引擎、提高财务预警准确度、识别上市公司年报错误以及预测企业内部控制重大缺陷，同时指出当前机器学习技术在我国智能财务运用中存在的问题[8]。

梁创维《基于机器学习的上市公司财务困境预警研究》中采用包裹式的指标选择方法，将预警指标筛选与随机森林建模过程相结合，对财务指标进行约简。对现有前沿的企业财务困境预警模型进行评价，选择一些现阶段学者重点关注的预警方法。最后，选择预警准确性最高的随机森林模型作为被解释模型，建立一个局部代理模型对其进行解释[9]。

3. 数据集

本文利用 Kaggle 平台分享的数据作为数据集，对企业员工流失进行分析。本文采用机器学习算法对数据进行模型构建和分析，并将 3 种算法模型(决策树模型、GBDT 模型、XGBoost 模型)进行对比，找出其中影响员工流失的主要因素，从而预测员工选择离职的各种可能性，并对企业提出建议和改进措施，有效的减少员工流失，降低企业成本。

3.1. 数据特征

本文根据 Kaggle 平台分享的数据集 HR_comma_sep.csv，设置相应的特征变量和目标变量。该数据集共有 14999 条记录，10 个变量，变量包括员工满意度(satisfaction_level)、最新绩效考核(last_evaluation)、参与项目数(number_project)、平均每月工作时长(average_montly_hours)、工作年限(time_spend_company)、是否发生过工作差错(Work_accident)、5 年内是否升职(promotion_last_5years)、部门(sales)、薪资(salary)等 9 个特征变量，以及 1 个目标变量，即是否离职(left)，测量类型、变量取值个数等如表 1 所示。

Table 1. Data characteristics
表 1. 数据特征

变量	不同取值个数	类型
员工满意度(satisfaction_level)	92	连续
最新绩效考核(last_evaluation)	65	连续
参与项目数(number_project)	6	分类
平均每月工作时长(average_montly_hours)	215	连续

Continued

工作年限(time_spend_company)	8	分类
是否发生过工作差错(Work_accident)	2	分类
5 年内是否升职(promotion_last_5years)	2	分类
部门(sales)	10	分类
薪资(salary)	3	分类
是否离职(left)	2	分类

3.2. 数据预处理

薪资(salary)为定序变量，本文将其取值字符转化为数值型，其中，“0”表示低水平，“1”表示中等水平，“2”表示高等水平。部门(sales)是定类型变量，对其进行 one-hot 编码。部门(sales)取值分别为 IT、RandD、accounting、hr、management、marketing、product_mng、sales、support、technical，哑变量处理后，生成 10 个哑变量，如表 2 所示。图 1 是对数据集中前 5 个员工部门特征取值的哑变量处理结果，从图 1 中可以看出，这 5 个员工都是销售部门的。

Table 2. Dummy variable handling
表 2. 哑变量处理

哑变量	部门									
	IT 部门	研发部门	会计部门	人事部门	管理部门	市场部门	产品部门	销售部门	运行部门	技术部门
1 sales_IT	1	0	0	0	0	0	0	0	0	0
2 sales_RandD	0	1	0	0	0	0	0	0	0	0
3 sales_accounting	0	0	1	0	0	0	0	0	0	0
4 sales_hr	0	0	0	1	0	0	0	0	0	0
5 sales_management	0	0	0	0	1	0	0	0	0	0
6 sales_marketing	0	0	0	0	0	1	0	0	0	0
7 sales_product_mng	0	0	0	0	0	0	1	0	0	0
8 sales_sales	0	0	0	0	0	0	0	1	0	0
9 sales_support	0	0	0	0	0	0	0	0	1	0
10 sales_technical	0	0	0	0	0	0	0	0	0	1

salary	sales_IT	sales_RandD	sales_accounting	sales_hr	sales_management	sales_marketing	sales_product_mng	sales_sales	sales_support	sales_technical
0	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0	0

Figure 1. Dummy variable result analysis
图 1. 哑变量结果分析

4. 模型建立

首先，提取特征变量 X，包括员工满意度(satisfaction_level)、最新绩效考核(last_evaluation)、参与项

目数(number_project)、平均每月工作时长(average_montly_hours)、工作年限(time_spend_company)、是否发生过工作差错(Work_accident)、5 年内是否升职(promotion_last_5years)、部门(sales)、薪资(salary)；提取目标变量 y，即员工是否离职(left)。其次，为避免出现数据过拟合的情况，将提取的数据划分成训练集(train)和测试集(text)两个部分，在训练集上建立模型(model)，然后再用测试集中的数据测试模型(model)分类预测效果。

4.1. 决策树模型

4.1.1. 决策树算法简介

决策树算法是一种典型的分类方法，首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策树对新数据进行分析。决策树本质上是通过一系列规则对数据进行分类的过程。决策树构造可以分两步进行。第一步，决策树的生成：由训练样本集生成决策树的过程。第二步，决策树的剪枝。决策树的剪枝是对上一阶段生成的决策树进行检验、校正和修剪的过程，主要是用新的样本数据集中的数据校验决策树生成过程中产生的初步规则，将那些影响预测准确性的分枝剪除。

4.1.2. 决策树模型搭建

图 2 为决策树测试集预测结果(Ddecisiontree test set prediction result)，将测试集中的前 20 个员工的 9 个特征数据代入建立好的决策树模型中，得到员工是否离职(left)的预测结果(0：表示未离职；1：表示离职)。图 2 中，预测离职人数为 6 人，但实际离职人数为 4 人。

sales_IT	sales_RandD	sales_accounting	sales_hr	sales_management	sales_marketing	sales_product_mng	sales_sales	sales_support	sales_technical	left	预测值
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	0	0	1	1
0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	1	1
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	1	1
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0

Figure 2. Decision tree test set prediction results
图 2. 决策树测试集预测结果

4.1.3. 特征变量重要性评估

在搭建模型时，提取的 9 个特征变量对员工是否离职(left)的影响是不一样的。图 3 给出了搭建决策树模型时不同特征变量对预测结果影响重要性的排序。特征变量对应的数值越高则表明该特征变量对员

工是否离职的影响越大。如图 3 所示，在决策树模型中，对员工是否离职(left)产生影响最大的是员工满意度(satisfaction_level)，其比重接近 9 个特征变量产生的影响总量的一半，而最新绩效考核(last_evaluation)、工作年限(time_spend_company)、参与项目数(number_project)、平均每月工作时长(average_montly_hours)这 4 个变量对员工是否离职(left)产生的影响效果中等，且这 4 个变量对结果的影响程度逐级递减。其余的薪资(salary)、部门(sales)、5 年内是否升职(promotion_last_5years)、是否发生过工作差错(Work_accident)对员工是否离职(left)产生的影响很小。

	特征名称	特征重要性
0	satisfaction_level	0.499112
1	last_evaluation	0.150738
4	time_spend_company	0.136254
2	number_project	0.103699
3	average_montly_hours	0.090045
7	salary	0.005536
16	sales_support	0.003435
17	sales_technical	0.003218
9	sales_RandD	0.001626
8	sales_IT	0.001604
15	sales_sales	0.000973
12	sales_management	0.000917
14	sales_product_mng	0.000839
13	sales_marketing	0.000707
11	sales_hr	0.000613
10	sales_accounting	0.000312
6	promotion_last_5years	0.000278
5	Work_accident	0.000094

Figure 3. Importance ranking of characteristic variables in decision tree model
图 3. 决策树模型特征变量重要性的排序

4.2. GBDT 模型

4.2.1. GBDT 算法简介

梯度提升决策树(Gradient Boosting Decision Tree, GBDT)算法是近年来被提及比较多的一个算法，是一种采用加法模型(即基函数的线性组合)与前向分步算法并以决策树作为基函数的提升方法。GBDT 是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论累加起来作为最终结果。GBDT 主要由 Decistion Tree (即 DT)、Gradient Boosting (即 GB)和 Shrinkage (步长)三部分组成。

4.2.2. GBDT 模型搭建

图 4 为 GBDT 模型测试集预测结果(GBDT model test set prediction result)，将测试集中的前 20 个员工的 9 个特征变量数据代入建立好的 GBDT 模型中，得到员工是否离职(left)的预测结果(0：表示未离职；1：表示离职)。图 4 中，预测离职人数为 4 人，实际离职人数也为 4 人。

sales_IT	sales_RandD	sales_accounting	sales_hr	sales_management	sales_marketing	sales_product_mng	sales_sales	sales_support	sales_technical	left	预测值
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	0	0	1	1
0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	1	1
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	1	1
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0

Figure 4. GBDT model test set prediction results
图 4. GBDT 模型测试集预测结果

4.2.3. 特征重要性评估

图 5 给出了搭建 GBDT 模型时不同特征变量对预测结果影响重要性的排序。如图 5 所示，在 GBDT 模型中，员工满意度(satisfaction_level)对员工是否离职(left)这一预测结果影响最大，其对预测结果的影响占比超过全部特征变量对结果影响的一半。工作年限(time_spend_company)、参与项目数(number_project)、最新绩效考核(last_evaluation)、平均每月工作时长(average_montly_hours)这 4 个特征变量对员工是否离职(left)的影响效果一般，且这 4 个变量对结果的影响程度逐级递减。薪资(salary)、是否发生工作差错(Work_accident)、部门(sales)这 3 个特征变量的重要性数值几乎为零，即这 3 个变量对员工是否离职(left)产生的影响极低，可忽略不计。另外，5 年内是否升职(promotion_last_5years)这一特征变量的重要性数值为 0，即 5 年内是否升职对最终的结果几乎没有影响。

4.3. XGBoost 模型

4.3.1. XGBoost 算法简介

XGBoost 是对梯度提升算法的改进，求解损失函数极值时使用了牛顿法，将损失函数泰勒展开到二阶，另外损失函数中加入了正则化项。XGBoost 在原有的 GBDT 基础上进行了改进，使得模型效果得到大大提升。作为一种前向加法模型，其核心是采用集成思想——Boosting 思想，将多个弱学习器通过一定的方法整合为一个强学习器，即用多棵树共同决策，每棵树的结果都是目标值与之前所有树的预测结果之差，并将所有的结果累加从而得到最终结果，以此达到整个模型效果的提升。

4.3.2. XGBoost 模型搭建

图 6 为 XGBoost 模型测试集预测结果(XGBoost model test set prediction result)，将测试集中的前 20 个员工的 9 个特征数据代入建立好的 XGBoost 模型中，得到员工是否离职(left)的预测结果(0：表示未离职；1：表示离职)。图 6 中，预测离职人数为 4 人，实际离职人数也为 4 人。

	特征名称	特征重要性
0	satisfaction_level	0.545646
4	time_spend_company	0.172701
2	number_project	0.108942
1	last_evaluation	0.104416
3	average_montly_hours	0.064698
7	salary	0.001759
5	Work_accident	0.001621
17	sales_technical	0.000089
10	sales_accounting	0.000072
14	sales_product_mng	0.000054
9	sales_RandD	0.000002
8	sales_IT	0.000000
6	promotion_last_5years	0.000000
11	sales_hr	0.000000
12	sales_management	0.000000
13	sales_marketing	0.000000
15	sales_sales	0.000000
16	sales_support	0.000000

Figure 5. Importance ranking of characteristic variables in GBDT model
图 5. GBDT 模型特征变量重要性排序

sales_IT	sales_RandD	sales_accounting	sales_hr	sales_management	sales_marketing	sales_product_mng	sales_sales	sales_support	sales_technical	left	预测值
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	0	0	1	1
0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	1	1
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	1	1
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0

Figure 6. XGBoost model test set prediction results
图 6. XGBoost 模型测试集预测结果

4.3.3. 特征变量重要性评估

图 7 给出了搭建 XGBoost 模型时不同特征变量对预测结果重要性的排序。如图 7 所示，在 XGBoost 模型中，员工满意度(satisfaction_level)、工作年限(time_spend_company)、参与项目数(number_project)这 3 个特征变量对员工是否离职(left)产生的影响程度最大，且这 3 个变量对测试结果的影响程度逐级递减。最新绩效考核(last_evaluation)、是否发生过工作差错(Work_accident)、平均每月工作时长(average_montly_hours)、薪资(salary)、部门(sales)、5 年内是否升职(promotion_last_5years)这 6 个特征变量对员工是否离职(left)产生的影响不大。

	特征名称	特征重要性
0	satisfaction_level	0.262601
4	time_spend_company	0.217166
2	number_project	0.135003
1	last_evaluation	0.086275
5	Work_accident	0.051208
3	average_montly_hours	0.040507
7	salary	0.025676
14	sales_product_mng	0.024068
11	sales_hr	0.021602
8	sales_IT	0.020677
10	sales_accounting	0.019784
17	sales_technical	0.018226
16	sales_support	0.016409
12	sales_management	0.014992
15	sales_sales	0.014043
6	promotion_last_5years	0.011924
9	sales_RandD	0.010376
13	sales_marketing	0.009464

Figure 7. Importance ranking of characteristic variables in XGBoost model
图 7. XGBoost 模型测试集预测结果

5. 模型的比较

5.1. 评价指标对比(准确率、精确率、召回率、F1 值和 CK 系数)

本文总共采用决策树算法、GBDT 算法和 XGBoost 算法等三种不同的算法，来预测员工是否离职。为了有效判断不同预测模型的性能，结合真实值，计算出准确率、精确率、召回率、F1 值和 Cohen’s Kappa 系数等评价指标，如表 3 所示。并根据表 3 中的数据绘制折线图(图 8)，对比 3 个模型的准确率、精确率、召回率、F1 值和 Cohen’s Kappa 系数，更好、更准确的评估这三个模型的性能。

Table 3. Model evaluation index comparison
表 3. 模型评价指标比对

模型 / 评价指标	决策树模型	GBDT 模型	XGBoost 模型
准确率	0.982	0.9756666666666667	0.99
精确率	0.9492455418381345	0.9635568513119533	0.9843081312410842
召回率	0.9760225669957687	0.9322990126939351	0.9732016925246827
F1 值	0.9624478442280946	0.9476702508960574	0.9787234042553192
Cohen’s Kappa 系数	0.9506139418671222	0.9318235306569971	0.9721877315757628

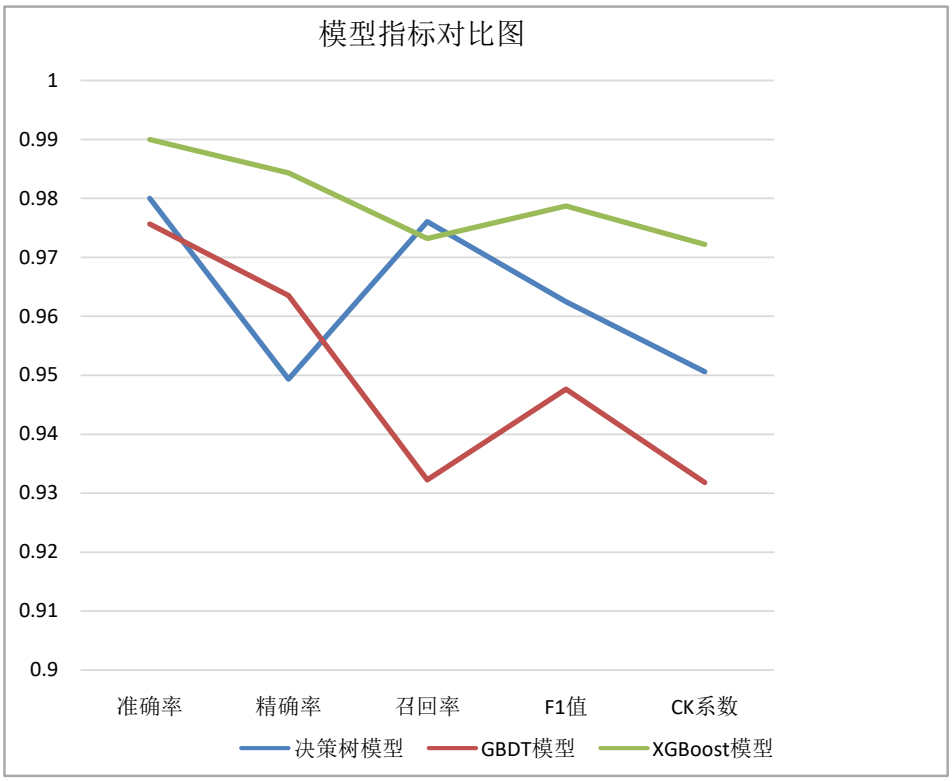

Figure 8. Model index comparison line chart
图 8. 模型指标对比折线图

准确率越高，预测正确的样本越多，模型的预测结果越准确，故准确率越大越好。由图 8 可看出三种模型中 XGBoost 模型的准确率最高，GBDT 模型的准确率最低；精确率越高则正确预测为“正”的样本越多，模型的预测结果越精确，故精确率越大越好。由图 8 可看出三种模型中 XGBoost 模型的精确率最高，决策树模型的精确率最低；召回率是实际为正的样本中被预测为正样本的概率，所以召回率越高越好。由图 8 可看出三种模型中决策树模型的召回率最高，GBDT 模型的召回率最低；F1 值越大则模型的质量越好，所以 F1 值越大越好。由图 8 可看出三种模型中 XGBoost 模型的 F1 值最大，GBDT 模型的 F1 值最小。Cohen’s Kappa 系数越大则一致性越高，所以 Cohen’s Kappa 系数越大越好。由折线图可看出三种模型中 XGBoost 模型的 Cohen’s Kappa 系数最大，GBDT 模型的 Cohen’s Kappa 系数最小。

综上，从图 8 中可以很清晰地看出这三种模型的性能，三种模型都能取得比较好的预测效果，其中，XGBoost 模型的性能最好，决策树模型的性能次之，GBDT 模型的性能相对较差。

5.2. ROC 曲线对比

在指标对比中，我们对三个模型的准确率、精确率、召回率、F1 值和 Cohen’s Kappa 系数五个指标进行了详细的对比。其中值得注意的是精确率和召回率互相影响，理想状态下肯定追求两个都高，但是实际情况是两者相互“制约”：追求精确率高，则召回率就低；追求召回率高，则通常会影响精确率。我们当然希望预测的结果精确率越高越好，召回率越高越好，但事实上这两者在某些情况下是矛盾的。这样就需要综合考虑它们。除了使用数值、表格形式评价分类模型的性能，也可以绘制出 ROC 曲线图，观察它们的分布情况。如图(图 9、图 10、图 11)所示，分别为决策树模型、GBDT 模型和 XGBoost 模型的 ROC 曲线。

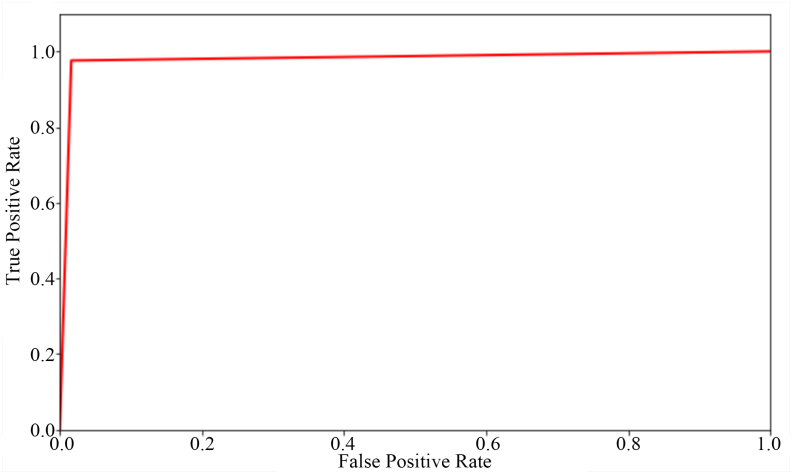

Figure 9. ROC curve of decision tree model
图 9. 决策树模型的 ROC 曲线

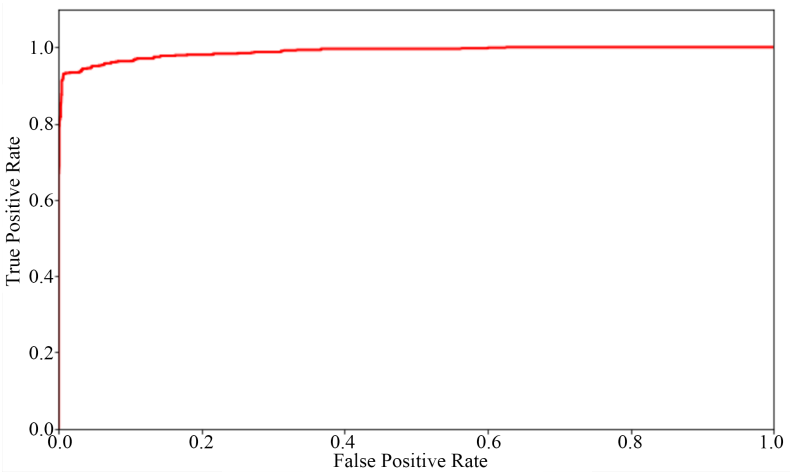

Figure 10. ROC curve of GBDT model
图 10. GBDT 模型的 ROC 曲线

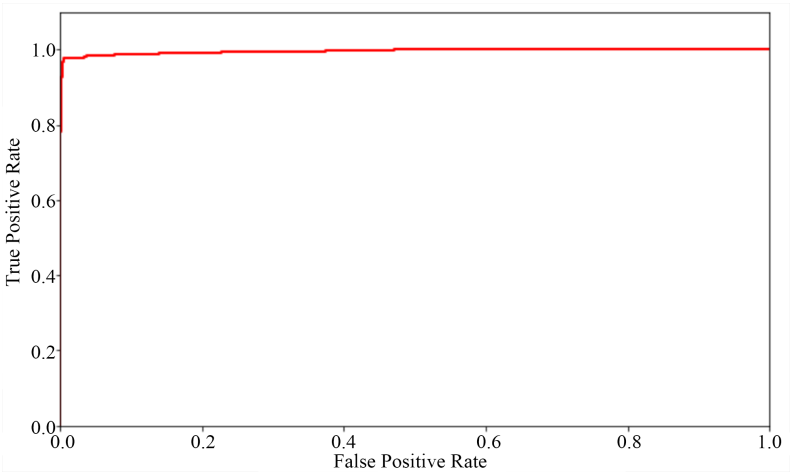

Figure 11. ROC curve of XGBoost model
图 11. XGBoost 模型的 ROC 曲线

在 ROC 曲线中，以假报警率(False Positive Rate)作为横坐标轴，命中率(True Positive Rate)作为纵坐标轴。AUC 值就是 ROC 曲线与 X 轴围成的图形面积，AUC 值越接近于 1 时模型性能最好。

Table 4. AUC value of three models
表 4. 三个模型的 AUC 值

模型 / AUC 值	决策树模型	GBDT 模型	XGBoost 模型
AUC 值	0.9799362071120267	0.9885690557088849	0.9955421318103156

如表 4，三种模型中 XGBoost 模型的 AUC 值最大，最接近于 1，故 XGBoost 模型的预测效果最准确。GBDT 模型的 AUC 值次之，故 GBDT 模型的预测效果次之。决策树模型的 AUC 值最小，故决策树模型的预测效果最差。综上，三个模型中预测效果最好的是 XGBoost 模型。

6. 总结

企业培养人才需要大量的成本，为了防止人才过度流失，企业应该注重员工流失分析，其员工流失分析是评估公司员工流动率的过程。本文基于 Kaggle 平台分享的数据集，在数据探索性分析和预处理的基础上，采用多种机器学习算法，构建企业员工流失预警模型，并进行模型的比较评价，目的是找到影响员工流失的主要因素，预测未来的员工离职状况，减少重要价值员工流失情况。

不过即使我们运用了多个算法模型，尽力控制好变量得出比较完美的结果，依然会发现本项目有一些不尽人意的地方，但这些归根到底来源于模型和算法本身的缺陷，例如：

1) 在决策树模型中，我们无法预测连续性的字段；由于员工离职的因素过多，即数据类别较多，决策树出现错误的概率也会相应增加，这也是三种模型对比时，决策树数据精确度最低的一个原因；同时，在处理员工流失这种特征关联度比较强的数据时，表现比较差。

2) 在 GBDT 算法中，虽然对企业员工流失的预测计算速度快，数据也较决策树精确率高，但是我们在实际运行过程中，需要仔细调参，花费了大量时间，同时在训练阶段，由于其是 boosting，存在串行结构，因此运行速度较慢。

3) 在 XGBoost 算法中，是三种模型中数据最优的一种，精确率和准确率都接近于 1，但是也存在局限性。由于对员工离职的数据较多，调参过程也很复杂，需要对 XGBoost 原理十分清楚才能很好的使用；如果将数据结果用图像的模式呈现，XGBoost 便不太适用，无法处理较高维度的特征数据。这也是之后本项目的一个拓展方向，即对比更多机器学习算法，选择较优的一种，对企业员工流失进行更加精确和准确的预测。因为存在这些缺陷，下一步我们将会参照更多的机器学习算法，构建模型，进行参数和结果对比，选取更具代表性的算法，利用其之间的优劣互补，更好地预测企业员工离职的影响因素，为企业员工流失预警提供决策支持。

另外，流失预警等算法模型终究是辅助性的，企业想要减少员工流失，最终是要靠企业自身的发展和规划，注重企业文化培养、员工职业生涯开发管理、凝聚企业向心力等内在措施，只有多管齐下，企业才能够获得更长远的发展。

基金项目

天津商业大学大学生创新创业训练计划市级项目，项目名称：基于机器学习的企业员工流失预警分析，项目编号：202110069104，项目类型：创新训练，项目领域：现代管理类。

参考文献

[1] 赵敏, 谭腾飞. 网络水军的成因及其发展——以库尔特·勒温“B=f(P·E)”为视角[J]. 新疆社科论坛, 2012(3): 64-66.

[2] Abdien, M.K. (2019) Impact of Communication Satisfaction and Work-Life Balance on Employee Turnover Intention. *Journal of Tourism Theory and Research*, **2**, 228-238. https://doi.org/10.24288/jttr.526678

[3] 张艳菊, 孙萌. 员工工作压力对员工离职倾向的影响模型构建[J]. 中外企业文化, 2020(10): 180-181.

[4] 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(3): 96-98.

[5] 王梦. 基于支持向量机的人力资源管理风险预警研究[J]. 化工管理, 2021(24): 3-4. https://doi.org/10.19900/j.cnki.ISSN1008-4800.2021.24.002

[6] 刘春燕. 基于 XGBoost 的员工流失预测研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2021. https://doi.org/10.26991/d.cnki.gdllu.2021.001604

[7] 杨守斌. 基于机器学习方法的 A 公司软件工程师绩效评价研究[D]: [硕士学位论文]. 青岛: 青岛科技大学, 2020. https://doi.org/10.27264/d.cnki.gqdhc.2020.000264

[8] 王玲. 机器学习技术在企业智能财务中的应用研究[J]. 商场现代化, 2022(13): 181-183. https://doi.org/10.14013/j.cnki.scxdh.2022.13.053

[9] 梁创维. 基于机器学习的上市公司财务困境预警研究[D]: [硕士学位论文]. 郑州: 中原工学院, 2021. https://doi.org/10.27774/d.cnki.gzygx.2021.000146