Computer Science and Application 计算机科学与应用, 2024, 14(11), 218-225
Published Online November 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.1411231

Hans汉斯

基于特征选择和机器学习的员工离职预测方法

罗洋雪绮[1]，何　林[2]

[1]重庆科技大学数理科学学院，重庆
[2]东北电力大学电气工程学院，吉林　吉林

收稿日期：2024年10月25日；录用日期：2024年11月22日；发布日期：2024年11月29日

摘　要

在大数据时代背景下，企业如何利用大数据技术及时掌握员工职业动态，提前洞悉并预测员工的离职倾向，帮助人力资源团队更好地作出人才挽留或储备的应对策略，这对企业的发展至关重要。本文以Datacastle平台的员工离职数据集为研究对象，首先对数据进行预处理、相关性分析、特征构造、非数值数据和数值型数据的处理以及不平衡数据的处理等步骤，然后利用递归特征消除算法反复构建模型剔除特征，最后利用决策树模型、支持向量机模型、逻辑回归模型、XGBoost模型分别对员工离职倾向进行预测，并对各模型的预测结果进行对比分析。结果显示，将smote采样处理后的数据应用于XGBoost模型后，无论在预测的准确率、召回率还是AUC的表现上均优于其他三个模型，作为员工离职预测的分类模型效果最佳。以该模型计算各变量的重要性排序，并结合交叉统计图分析后得出，员工婚姻状况、所学习的专业领域、所在部门、股票期权水平等因素对员工是否离职的影响较高。

关键词

随机森林，数据处理，决策树模型，递归特征消除

Employee Turnover Prediction Method Based on Feature Selection and Machine Learning

Xueqi Luoyang[1], Lin He[2]

[1]School of Mathematics and Big Data, Chongqing University of Science and Technology, Chongqing
[2]School of Electrical Engineering, Northeast Electric Power University, Jilin Jilin

Received: Oct. 25th, 2024; accepted: Nov. 22nd, 2024; published: Nov. 29th, 2024

文章引用: 罗洋雪绮, 何林. 基于特征选择和机器学习的员工离职预测方法[J]. 计算机科学与应用, 2024, 14(11): 218-225. DOI: 10.12677/csa.2024.1411231

Abstract

In the era of digital economy, if enterprises make full use of science and technology to predict the turnover dynamics of employees and understand the turnover tendency of employees in advance, it can help the human resources team to make better coping strategies of talent retention or reserve, which has far-reaching significance for the development of enterprises. The results show that applying SMOTE-sampled data to the XGBoost model outperforms the other three models in terms of prediction accuracy, recall rate, and AUC. It proves to be the most effective classification model for employee turnover prediction. Based on the importance ranking of variables calculated by this model and analyzed with cross-statistical charts, it was found that factors such as the employee's marital status, field of study, department, and stock option level have a significant impact on whether the employee leaves the company.

Keywords

Random Forest, Data Processing, Decision Tree Model, Recursive Feature Elimination

1. 背景介绍

习近平总书记指出，人才是推动企业发展的关键，必须实施人才强国战略。在科技迅速进步的背景下，人才成为企业持续成长的核心。尽管如此，企业正面临人才留存的挑战。前程无忧的报告揭示了 2023 年员工离职率为 17.9%。员工流失不仅意味着人员离开，还会给企业带来多方面的影响。因此，人力资源部门需提前了解员工离职原因，采取措施如改善工作环境、提升薪酬和职业发展机会，以降低流失率。随着机器学习和数据挖掘技术在企业管理中的应用，研究员工流失问题变得越来越重要。本文通过机器学习模型分析员工离职数据，旨在找出预测效果最佳的模型，并识别主要离职因素，以期为人力资源政策提供参考。

2. 相关文献综述

离职问题中的传统机器学习理论的应用研究。Sisodia 在其研究中，基于人力资源数据集，采用了多种机器学习算法，包括支持向量机(SVM)、C5.0 决策树、随机森林以及 k 最近邻等，构建了员工离职预测模型[1]。通过这些模型，他深入剖析了员工离职的各种潜在因素，并指出了企业需要针对这些因素进行优化，以降低员工离职率，提升企业的稳定性和竞争力。李芸在其研究中，针对国网电力公司的员工离职问题进行了深入探讨。她巧妙地运用了员工离职数据集，结合支持向量机(SVM)的多种核函数，成功构建了员工离职预测模型[2]。

离职问题中数据不平衡问题的相关研究。Gao 在深入研究员工离职问题的基础上，通过改进模型，提出了一种改进的随机森林算法，该算法专门用于处理不平衡的员工离职数据。相较于传统的随机森林算法，这种改进版在应对离职数据不平衡时表现出色，其计算精度也得到了显著提升[3]。在实际应用中，这种算法有效地帮助企业和组织识别离职风险，提前采取措施以减少人才流失。万毅斌则针对员工离职数据中的不平衡问题，提出了一种改进的自适应模糊 C 均值聚类算法[4]。该算法结合了 SMOTE 采样算

法和 SVM 算法的核技巧，提出了一种基于核空间的聚类过采样算法。这种算法不仅能够有效地处理不平衡数据，还能提高分类器的泛化能力。在此基础上，他进一步结合集成学习算法，提出了针对不同类型企业的综合离职预测模型。这种模型能够根据不同企业的特点和需求，提供个性化的离职预测方案，为企业的人力资源管理提供有力支持。

离职问题中数据可视化的相关研究。为更深入地了解员工离职问题的本质，从而为企业制定更有效的人力资源管理策略提供有力的支持，有学者通过数据可视化的方法直观展现相关的结果。陆亚琳深入研究了数据可视化技术在特征选择方面的应用。她首先通过精心挑选和整理大量的员工离职数据，运用数据可视化的方法，将这些数据以直观、易于理解的形式展现出来[5]。这种可视化技术不仅有助于快速识别离职员工的关键特征，还能够帮助企业更好地理解员工离职的潜在原因。在特征选择的基础上，陆亚琳进一步利用朴素贝叶斯和决策树这两种经典的机器学习算法来构建员工离职预测模型。她通过对比不同模型在离职预测方面的性能，发现朴素贝叶斯和决策树模型在预测员工离职方面具有较高的准确性。此外，她还从多个角度对员工离职预测模型进行了深入的分析与研究，包括模型的稳定性、可解释性以及在实际应用中的可行性等方面。Wild 利用主成分分析对员工进行特征选择，通过降维技术将原始的高维数据转化为低维的主成分变量，从而更容易地揭示数据中的关键信息。在此基础上，他使用随机森林算法建立了员工离职模型，并通过实验验证了该模型的准确性。Wild 的研究结果表明，基于主成分分析法的随机森林模型在员工离职预测方面具有更高的准确度[6]。这种模型不仅能够有效地识别出影响员工离职的关键因素，还能够为企业提供有针对性的建议，帮助企业更好地应对员工离职问题。

离职问题中各类机器学习混合模型的应用研究。Yunmeng 在其研究中，巧妙地运用了 K-Means 算法对员工进行了细致的分类，随后使用决策树模型进行离职预测分析[7]。这一方法不仅提高了预测的准确性，还有助于企业更好地理解员工离职的潜在原因，从而制定出更为有效的员工离职管理策略。Yogesh 通过运用多种模型进行了员工离职预测的研究，他使用了决策树、AdaBoost 和随机森林等模型，通过比较不同模型的预测效果，为企业提供了更加全面的离职预测参考[8]。这些模型各自具有不同的特点和优势，能够从不同的角度揭示员工离职的复杂因素。此外，李强则在其研究中采用了更为先进的模型融合技术。他利用 AdaBoost 和 Random Forest 模型作为基础模型，通过 Stacking 模型融合的方式，构建了 LRA 员工离职预测模型。这种模型融合技术能够充分利用各个基础模型的优点，提高整体预测性能。同时 LRA 模型还能够帮助企业更准确地评估员工离职风险，为企业制定人力资源管理策略提供有力支撑[9]。

3. 数据处理

(一) 数据来源

本文使用的是 Github 平台中的员工离职预测数据集，来源于 IBM Waton Analytics 分析平台分享的真实数据。平台提供的训练数据主要包括 1471 条记录，共 38 个属性，个别数据的属性具体说明如表 1 所示。

Table 1. Summary table of data attributes
表 1. 数据属性汇总表

序号	字段名	含义	说明
1	Age	员工年龄	数值
2	Attrition	员工离职状态	其中，0 表示已离职，1 表示未离职，这是目标预测值，即目标属性
3	Business Travel	商务差旅频率	Travel Rarely 表示不经常出差 Travel Frequently 表示经常出差 Non Travel 表示不出差

续表

4	Department	员工所在部门	Sales 表示销售部 Research & Development 表示研发部 Human Resources 表示人力资源部
5	Distance From Home	公司跟家庭住址的距离	从 1 到 29，1 表示最近，29 表示最远
6	Education	员工的受教育程度	从 1 到 5，5 表示教育程度最高
7	Eduration Field	员工所学习的专业领域	Life Sciences 表示生命科学 Technical Degree 表示技术学位 Human Resources 表示人力资源 Other 表示其他
8	Environment Satisfaction	员工对于工作环境的满意程度	从 1 到 4，1 的满意程度最低，4 的满意程度最高
9	Job Role	工作角色	Sales Executive 是销售主管 Research Scientist 是科学研究员 Laboratory Technician 是实验室技术员 Manufacturing Director 是制造总监 Healthcare Representative 是医疗代表 Manager 是经理 Sales Representative 是销售代表 Research Director 是研究总监 Human Resources 是人力资源
10	JobSsatisfaction	工作满意度	从 1 到 4，1 代表满意程度最低，4 代表满意程度最高
11	Marital Status	员工婚姻状况	Single 代表单身 Married 代表已婚 Divorced 代表离婚
12	Monthly Income	员工月收入	数值
13	Num Companies Worked	员工曾经工作过的公司数	数值
14	Percent Salary Hike	工资提高的百分比	数值

(二) 探索性分析

在该数据集中，拥有 38 个属性，涵盖了员工的多方面信息。这些属性中，最重要的是 Attrition，这是目标属性，它揭示了员工的离职状态。Attrition 的研究对于公司来说至关重要，因为它直接关系到公司的员工稳定性、工作效率和整体竞争力。

由于数据集中的“Employee Number”只是员工的唯一标识符，没有提供关于员工行为或态度的任何有价值的信息。因此在后续分析中删除这一属性。在剩余的属性中，月费率(MonthlyRate)、小时费率(HourlyRate)、日费率(DailyRate)这三个属性在本质上都是衡量员工薪酬的指标，只是时间单位不同。为了避免数据冗余和共线性，本文主要对月收入(MonthlyRate)进行探索性分析，具体属性见表 1。

数据分为三类：目标属性、连续属性和分类属性。文章首先用扇形图展示目标属性分布，了解员工离职比例。接着，用簇状柱形图比较不同月收入员工的离职率，识别高离职风险收入段。此外，使用百分比堆积图分析目标属性与连续、分类属性的关系，探究影响离职率的因素。通过这些分析，全面理解员工离职原因，为员工管理和薪酬策略提供决策支持，并指导数据预处理。

(三) 数据清理

数据清理是数据科学中确保模型性能的关键步骤，直接影响预测结果。我们用 Jupyter notebook 展示

数据清理的有效方法。首先，识别并去除噪声和冗余信息，提高模型效率和准确性。其次，用 Pearson 相关系数法过滤数据，降低模型复杂度，增强泛化能力。最后，处理非数值型数据、数值型数据和不平衡数据，确保数据质量，为数据分析和模型训练打下坚实基础，构建性能优秀的机器学习模型。

4. 员工离职预测模型

(一) 特征选择

特征选择在机器学习中至关重要，它帮助模型识别并保留对预测结果有影响的特征，同时排除无关或冗余的特征，以优化模型性能。通过减少特征数量，特征选择可以降低模型复杂度，防止过拟合，提高泛化能力，并使模型更简洁易懂。此外，它还能提升训练速度，特别是在处理大规模数据集时。特征选择方法多样，包括统计、模型和机器学习方法。本项目采用递归特征消除(RFE)，它通过递归减少特征集规模来评估特征重要性，适用于能够提供特征重要性排序的学习算法，如 SVM 和随机森林。RFE 特别适用于特征数量多且存在多重共线性的情况，有助于消除冗余特征，增强模型泛化能力。

(二) 模型实现

机器学习算法是人工智能的关键部分，包含多种模型，各有其优势和应用场景。员工离职预测是企业人力资源管理的核心任务，准确预测有助于减少人才流失。随着机器学习技术的进步，多种算法被用于此预测。本项目采用决策树、逻辑回归、支持向量机和 XGBoost 构建员工离职预测模型。

1) 决策树模型的建立

决策树是一种监督学习算法，通过划分数据集构建树形结构进行分类或预测。每个节点代表属性测试，分支代表测试结果，叶节点代表预测结果。构建过程涉及选择最优特征以提高数据“纯净度”。随机森林是集成学习方法，通过多个决策树预测，每棵树在随机数据子集和特征子集上独立训练。对比正常数据和 SMOTE 处理后的数据在决策树和随机森林模型中的性能，结果包括召回率、准确率和 Cohen’s Kappa 系数，见表 2。

Table 2. Decision tree model evaluation
表 2. 决策树模型评估

Model Name	Accuaracy	Kappa Score	ROC AUC Curve value
Decision tree with normal data	84.23%	0.25	0.61
Random forest with normal data	87.72%	0.26	0.59
Decision tree with smote data	79.61%	0.25	0.64
Random forest with smote data	86.68%	0.36	0.65

表 2 显示，随机森林在正常数据上更准确，在 SMOTE 数据上更可靠。这表明正常数据下随机森林性能更优，能有效学习信息。当数据不平衡时，SMOTE 技术提升算法处理能力，增强可靠性。

2) 逻辑回归模型

逻辑回归是一种用于二分类问题的经典模型，尽管名称包含“回归”，但它实际上是一种分类算法。该模型旨在根据输入特征对样本进行分类，通过特征的线性组合和一个非线性激活函数将结果转换为概率值。

逻辑回归模型的数学表示如下：

$$y = sigmoid(X * beta) \tag{1}$$

其中，y 表示预测的类别概率，X 是输入特征矩阵，$beta$ 是模型的参数。$sigmoid$ 函数可以将线性组合的结果映射到 0 到 1 之间的概率值，一般采用如下形式的 $sigmoid$ 函数：

$$sigmoid(z) = 1/(1+\exp(-z)) \tag{2}$$

逻辑回归模型通过最大化似然或最小化损失函数来确定参数 $beta$。常用的损失函数是交叉熵，用于评估模型预测与实际类别的差异。对比正常数据和 smote 处理数据在逻辑回归模型中的表现，结果包括准确率、Kappa 系数、召回率等，详见表 3。

Table 3. Logistic regression model evaluation
表 3. 逻辑回归模型评估

Model Name	Accuaracy	Kappa Score	ROC AUC Curve value
Logistic Regression with normal data	88.04%	0.42	0.67
Logistic Regression with smote data	78.53%	0.35	0.73

从表 3 中可以明显看出，在正常数据的情况下，逻辑回归模型表现更为出色。这一结果进一步证实了逻辑回归在处理分类问题时的有效性，尤其是在数据分布符合其假设条件时。

3) 支持向量机模型

支持向量机(SVM)是一种用于分类和回归的监督学习算法，旨在通过找到一个超平面来最大化不同类别间的间隔。在二分类问题中，SVM 通过最大化间隔来分隔两个类别，其中最接近超平面的数据点(支持向量)决定了超平面的位置。对于非线性问题，SVM 使用核技巧将数据映射到高维空间以实现线性可分。常见的核函数有线性核、多项式核和高斯核等。

为了训练高效的 SVM 模型，我们采取了严格的数据分割策略，将数据分为训练集、测试集和验证集，以评估模型表现并调整参数。在训练过程中，我们特别关注超参数 C 的选择，它影响模型的正则化程度和对误分类样本的惩罚。C 值较大时，模型会减少训练集中的错误分类，但可能导致过拟合；而较小的 C 值可能减少模型复杂度，避免过拟合，但可能导致欠拟合。

将不同的核函数和 C 值组成四个模型对比，确认最优模型后对模型进行训练并可视化特征的系数以及评估性能、计算 kappa 系数，得到的数据如表 4 所示：

Table 4. SVM model evaluation
表 4. 支持向量机模型评估

Model Name	Accuaracy	Kappa Score	ROC AUC Curve value
SVM	83.9%	0.19	0.58

在支持向量机模型中采用线性核函数，C 值为 0.01 效果最好。

4) XGBoost 模型

Table 5. XGBoost model evaluation
表 5. XGBoost 模型评估

Model Name	Accuaracy	Kappa Score	ROC AUC Curve value
XGBoost with normal data	88.04%	0.40	0.66
XGBoost with smote data	88.31%	0.44	0.68

XGBoost 通过添加新的决策树来提升模型性能，尤其擅长处理复杂非线性问题。但有时它可能不足以达到最佳精度，因此我们采用堆叠技术，结合多个模型的预测结果来训练一个最终模型。本项目中，我们将决策树模型叠加在 XGBoost 之上，利用决策树来学习并优化 XGBoost 的预测模式，从而捕捉更多数据细节。

将正常数据和 smote 过采样技术处理过的数据代入 XGBoost 模型对比，得到的准确率、Kappa 系数以及召回率、准确率的值如表 5 所示。

总的来说使用 smote 数据集得到的 XGBoost 模型更优。

5. 总结

本文针对员工离职问题，以 Github 平台上提供的真实员工离职数据集为研究对象，在对数据进行初步处理之后，我们进一步进行了相关性分析。通过分析发现，员工的工作满意度、晋升机会、工作压力等因素与离职意向之间存在显著的相关性。这些发现为我们后续的特征构造提供了重要的参考。本文针对非数值型和数值型数据分别进行了处理。对于非数值型数据，本文采用了 one-hot 编码的方式进行转换，以便后续模型的训练。对于数值型数据进行归一化处理，以消除不同特征之间的量纲差异对模型的影响。考虑到数据不平衡问题，本文采用了 smote 过采样技术进行处理。通过对少数类样本进行过采样成功地使数据达到了平衡状态，从而避免了数据不平衡对模型性能的影响。

在构建员工离职预测模型时，本文采用了递归特征消除方法对特征进行了筛选，去除了冗余和无关的特征，提高了模型的泛化能力。随后，分别构建了决策树模型、逻辑回归模型、支持向量机模型和 XGBoost 模型，并对各个模型的性能进行了对比。实验结果显示，经过 smote 过采样技术处理过的数据代入 XGBoost 模型得到的模型性能最好。

当在深入研究本文所提供的观点后，我们可以发现确实存在一些不足之处，这些不足在一定程度上影响了模型的预测能力和方法的可靠性。以下是对这些不足的深入分析和进一步的探讨。首先，调参的困难是机器学习项目中常见的挑战之一。对于本文所涉及的项目，由于涉及的参数较多，仅仅对支持向量机模型的 C 值进行调整是远远不够的。参数的优化是一个复杂而耗时的过程，需要对算法有深入的理解和实践经验。在未来的工作中，可以尝试使用网格搜索、随机搜索或贝叶斯优化等调参方法，对其他算法的参数进行更细致的调整，从而充分发挥模型的预测能力。其次，数据来源单一和数据维度不充分也是本文的一个显著问题。仅仅依赖 Github 平台上提供的单一数据集来评估方法的有效性是不够全面的。在实际应用中，不同的数据集可能会表现出不同的特征和规律，因此，使用多个数据集进行对比实验是必要的。此外，数据维度的不足也可能导致模型无法捕捉到足够的信息，从而影响预测结果的准确性和可靠性。为了改进这一点，可以考虑增加一些其他相关指标，以丰富数据维度，进而提高模型的性能和稳定性。

总之，虽然本文在某些方面表现出了一定的不足，但这些不足也为后续的研究提供了改进的方向和思路。通过更深入的研究和实践，相信可以不断优化模型和方法，提高预测结果的准确性和可靠性，为相关领域的发展做出更大的贡献。

基金项目

重庆市研究生科研创新项目(YKJCX2321101)。

参考文献

[1] Sisodia, D.S., Vishwakarma, S. and Pujahari, A. (2017) Evaluation of Machine Learning Models for Employee Churn Prediction. 2017 *International Conference on Inventive Computing and Informatics* (*ICICI*), Coimbatore, 23-24

November 2017, 1016-1020. https://doi.org/10.1109/icici.2017.8365293

[2] 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(6): 20-22.

[3] Gao, X., Wen, J. and Zhang, C. (2019) An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering*, **2019**, Article ID: 4140707. https://doi.org/10.1155/2019/4140707

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

[5] 陆亚琳. 数据挖掘技术在人力资源管理系统中的应用研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2020.

[6] Wild Ali, A.B. (2021) Prediction of Employee Turn over Using Random Forest Classifier with Intensive Optimized Pca Algorithm. *Wireless Personal Communications*, **119**, 3365-3382. https://doi.org/10.1007/s11277-021-08408-0

[7] Yunmeng, Z. and Chengyi, Z. (2019) The Application of the Decision Tree Algorithm Based on K-Means in Employee Turnover Prediction. *Journal of Physics*: *Conference Series*, **1325**, Article 012123. https://doi.org/10.1088/1742-6596/1325/1/012123

[8] Yogesh, I., Suresh Kumar, K.R., Candrashekaran, N., Reddy, D. and Sampath, H. (2020) Predicting Job Satisfaction and Employee Turnover Using Machine Learning. *Journal of Computational and Theoretical Nanoscience*, **17**, 4092-4097. https://doi.org/10.1166/jctn.2020.9024

[9] 李强, 翟亮.基于 Stacking 算法的员工离职预测分析与研究[J]. 重庆工商大学学报(自然科学版), 2019, 36(1): 117-123.