

基于随机森林算法的AI对大学生就业压力的影响分析

赵伊健, 刘大元, 顾俊杰

苏州工学院数学与统计学院, 江苏 苏州

收稿日期: 2025年4月7日; 录用日期: 2025年4月28日; 发布日期: 2025年5月12日

摘要

为了更好地分析AI对大学生就业带来压力, 本研究融合经典统计检验(T检验)与机器学习算法; 通过贝叶斯优化对随机森林超参数调优; 利用最佳参数组合对样本进行建模, 挖掘出AI给大学生带来就业压力的主要影响因素。模型评价指标如下: 准确率(0.8755), 精确率(0.8857), 召回率(0.9051), F1分数(0.8953), AUC-ROC值(0.9399), AUC-PR值(0.9441)。结果表明, 本研究的模型分类效果具有优越性。我们通过建立模型并分析得出AI的专业关联、AI行业与就业冲击接纳等指标因素对大学生的AI就业压力感知影响重大。我们的研究结果可以为高校应对AI时代的学科调整和大学生就业指导以及为大学生的自我提升与就业提供了相应的理论支持, 帮助大学生更好地应对AI时代的就业挑战。

关键词

T检验, 随机森林, 贝叶斯优化

Analysis of the Impact of AI on College Students' Employment Pressure Based on the Random Forest Algorithm

Yijian Zhao, Dayuan Liu, Junjie Gu

School of Mathematics and Statistics, Suzhou University of Technology, Suzhou Jiangsu

Received: Apr. 7th, 2025; accepted: Apr. 28th, 2025; published: May 12th, 2025

Abstract

In order to better analyze the pressure brought by AI on college students' employment, this study integrates classical statistical tests (t-test) and machine learning algorithms. The hyperparameters of the

文章引用: 赵伊健, 刘大元, 顾俊杰. 基于随机森林算法的 AI 对大学生就业压力的影响分析[J]. 统计学与应用, 2025, 14(5): 11-20. DOI: 10.12677/sa.2025.145121

random forest are tuned through Bayesian optimization. The optimal parameter combination is used to model the samples, and the main influencing factors of AI-induced employment pressure on college students are excavated. The model evaluation indicators are as follows: accuracy (0.8755), precision (0.8857), recall (0.8832), F1-score (0.8953), AUC-ROC value (0.9399), and AUC-PR value (0.9441). The results show that the classification effect of the model in this study is superior. Through model building and analysis, we have found that factors such as the correlation between majors and AI, and the acceptance of AI-related industry and employment impacts have a significant influence on college students' perception of AI-related employment pressure. Our research results can provide theoretical support for colleges and universities to cope with the discipline adjustment and employment guidance of college students in the AI era, as well as for the self-improvement and employment of college students, and help college students better cope with the employment challenges in the AI era.

Keywords

T-Test, Random Forest, Bayesian Optimization

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

电气时代的自动化程度提高,曾引起工人的恐慌,使其害怕工作岗位丢失,而近年来,人工智能的快速发展,使它在现实社会中的应用变得广泛,不免令人焦虑。而大学生又是就业市场关键且主要的新生力量,首当其冲会受到 AI 发展的影响。大学生们怀揣着对未来的憧憬步入校园,但是在毕业时又要面临着 AI 所影响的就业环境的变革。一方面,专业与 AI 高度相关的学生将会迎来前所未有的机遇,因为 AI 技术的广泛应用为他们创造了大量与之相关的新岗位;另一方面,所学专业与 AI 相关程度低的学生则可能面临传统岗位将会被 AI 替代的风险,并且就业竞争也愈发激烈。因此,深入研究 AI 对大学生就业压力的影响,成为当今社会亟待解决的重要课题。

传统问卷调查方法存在一定的局限性,且该领域的机器学习算法还未充分应用。在研究 AI 对大学生的就业压力影响时,存在着众多相互影响的因素,比如 AI 技术的发展、AI 应用领域的拓展和大学生个体之间的差异如专业背景、职业规划等因素,传统的问卷调查难以分析这些因素间的非线性关系与复杂交互作用,这将会导致研究结果与实际结果存在偏差。

为了解决上述问题,本研究以大学生对于 AI 带来的就业压力作为切入点展开调查分析,通过线上发布问卷收集数据,基于此调查数据建立了机器学习中的随机森林模型用于分类预测 AI 技术带来的就业压力感知,并深入分析其中的关键影响因素,为高校与大学生未来在 AI 时代的就业能力培养与就业方向调整提供理论支持。

2. 数据获取与预处理

2.1. 数据获取

本研究采用线上问卷调查的方式收集数据,借助专业问卷平台广泛发布问卷,主要的调查对象是在校大学生。问卷设计围绕大学生对 AI 的认知、态度、学习与应用情况等,以及 AI 对大学生就业压力的影响等方面进行展开,通过线上发布问卷收集数据,截止 2025 年 3 月 31 日共获得了 656 份问卷,并以 Excel 文件形式保存数据。

2.2. 预处理

基于研究理论框架, 首先, 对原始问卷指标进行重命名, 共获得如下指标: AI 好奇心, AI 学习意愿, AI 创新应用, 创造性思维, AI 知识了解 AI 应用举例, AI 专业关联, AI 应用接纳, AI 工作接纳, AI 就业冲击接纳, AI 行业冲击接纳, AI 就业压力等。其次, 根据理论构建新的交叉特征, 将知识_应用协同记为 K, AI 知识了解记为 N, AI 创新应用记为 I, 则:

$$K = N \times I$$

将分类变量进行特征映射编码, 并将指标 AI 就业压力分类, 1、2、3 为低压力组, 4、5 为高压力组, 用于后续分析。最后, 调整完毕后, 用于后续的特征工程。

2.3. 特征预筛选

为了进行特征预筛选, 我们对选取的特征使用 t 检验结合 Bonferroni 校正, 观察低压力组与高压力组之间有显著差异的指标, 组间关系显著的指标用于后续的模式训练。校正检验结果详见表 1。

Table 1. Calibration test result data
表 1. 校正检验结果数据

特征	t 统计量	效应量(Cohen's d)	校正后 p 值	是否显著
AI 好奇心	-6.703097	0.976509	1.271674e-08	显著
AI 就业冲击接纳	-6.613710	0.922599	1.853517e-08	显著
AI 专业关联	-6.583165	0.962923	2.284484e-08	显著
AI 学习意愿	-5.812941	0.807828	8.172074e-07	显著
AI 行业冲击接纳	-5.222733	0.711972	1.130829e-05	显著
AI 知识了解	-5.183545	0.746649	1.386546e-05	显著
AI 应用举例	-4.747110	0.683201	8.705882e-05	显著
知识_应用协同	-4.615661	0.666630	1.488480e-04	显著
AI 创新应用	-4.559970	0.638363	1.839702e-04	显著
AI 应用接纳	-4.498850	0.660772	2.390119e-04	显著
创造性思维	-4.374893	0.616010	3.838923e-04	显著
AI 工作接纳	-4.299576	0.638373	5.229967e-04	显著
年级	0.250385	-0.028935	1.000000e+00	不显著

从特征分析结果来看, 年级这个特征在低压力和高压力组之间无显著差异, 这可能是因为不同年级的大学生在面对 AI 造成的就业压力时, 受到的影响较为相似, 也可能是问卷中的年级划分不够细致, 未能体现出差异。对于其他特征, 如 AI 好奇心、AI 就业冲击力接纳以及 AI 学习意愿等, 在两组间差异显著, 表明这些因素与大学生对 AI 带来的就业压力感知情况密切相关, 为后续模型训练提供了重要依据。

3. 模型建立与分析

3.1. 模型建立

随机森林是一种集成学习分类器, 可通过构建多棵基于特征子集的决策树, 显著提高分类性能, 特别是在小规模训练集上表现出色[1]。其构建过程包括以下关键步骤: 1) 训练子集生成。通过自助采样法

(Bootstrap Sampling)从原始数据集中抽取多个训练子集。2) 决策树独立构建。对每个训练子集独立构建一棵决策树，在树的每个节点分裂时，通过不放回的随机抽样方式选取候选特征集，增加模型的多样性和鲁棒性。3) 特征选择与节点划分。依据基尼指数增益最大化原则，从候选特征中选取最优特征及相应的分割点进行节点划分，直至满足停止准则，确保决策树的有效生长。4) 预测结果集成。通过多数投票机制整合所有决策树的预测输出，得出随机森林的最终分类决策，此过程有效提升了模型的预测精度和稳定性。且为了保证模型的泛化能力，随机森林在生成每棵决策树时，往往会遵循数据随机和特征随机的基本原则[2]。具体步骤见下图 1。

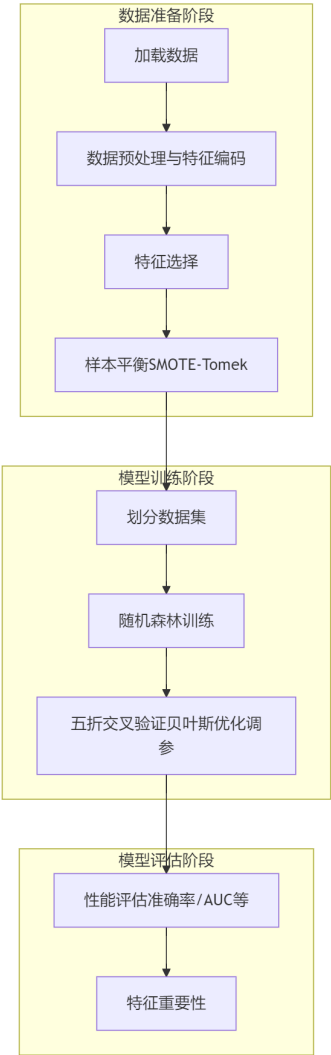


Figure 1. Random forest flowchart
图 1. 随机森林流程图

原始数据存在着类别不平衡，采取过采样和欠采样，将少数类采样比例设置为 0.7，解决类别相差大导致的模型分类效果不理想的情况。随机森林模型在处理高维特征、非线性关系的捕捉方面表现优异，因此本研究使用此模型并使用贝叶斯优化搜索最佳超参数组合，优化目标为 5 折交叉验证的 AUC-ROC 均值，交叉验证的操作可提高模型的稳定性。

在本研究中,采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值(F1-score)、受试者工作特征曲线下面积(AUC-ROC)和召回率曲线下面积(AUC-PR)等指标来评估就业压力分析模型的性能,这些指标在大学生就业压力分类任务中具有重要作用。TP 指在数据集中实际为正面且被准确识别为高压力的样本数量;FP 则表示实际为低压力却被误判为高压力的样本数量;FN 指低压力被误判为高压力的样本数量,而 TN 则是低压力被准确识别的样本数量。为了方便起见,在此,记准确率为 A、精确率为 P、召回率为 R、受试者工作特征曲线下面积为 AR,以及精确率-召回率曲线下面积为 AP。计算公式分别如下[3]:

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \tag{1}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{2}$$

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{3}$$

$$F_1 = \frac{2 \times A \times R}{A + R} \tag{4}$$

$$AR = \int_0^1 N_{TP} \times N_{FP} dN_{FP} \tag{5}$$

$$AP = \int_0^1 P \times R dR \tag{6}$$

通过对大学生各项就业高压力与低压力样本的分析,我们得到了关于评判高压力与低压力样本正误的混淆矩阵,结果见表 2:

Table 2. Confusion matrix
表 2. 混淆矩阵

	预测为正例	预测为反例
实际为正例	80	13
实际为反例	16	124

由表 2 的混淆矩阵得,模型在预测高压力样本时,有 80 个样本被正确预测,有 13 个样本被误判;在预测低压力样本时,有 16 个样本被误判,124 个样本被正确预测。对于混淆矩阵的数据,我们通过计算误分类率,来深入反映模型的性能。高压力样本的误分类率为 $13 \div (80 + 13) \approx 0.1398$,同理低压力样本的误分类率为 $16 \div (16 + 124) \approx 0.1143$ 。由此得出,虽然低压力样本误分类率略低于高压力样本,但是二者均处于一定的水平,这也说明了在经过选择的特征下该模型拥有良好的分类能力。

通过贝叶斯优化后的超参数组合输出的模型表现如下表 3。

由模型评估指标得,准确率达到 0.8755,可以得出模型整体的分类准确性较高;模型的精确率是 0.8857,说明模型在分类预测为高压力的样本中,有较高比例是真正的高压力样本;模型召回率为 0.9051 表示模型能够较好地识别出实际的高压力样本;F1 的分数为 0.8953,将精确率和召回率综合,进一步体现了模型在分类性能上的平衡;AUC-ROC 和 AUC-PR 的值分别为 0.9399 与 0.9441 都接近 1,说明模型在区分高压力和低压力样本方面表现优秀,具有良好的分类性能。

Table 3. Model evaluation metrics
表 3. 模型评估指标

指标	值
准确率	0.8755
精确率	0.8857
召回率	0.9051
F1 分数	0.8953
受试者工作特征曲线下面积	0.9399
召回率曲线下面积	0.9441

3.2. 模型指标评价

通过 python 程序对多个模型指标进行可视化，从图 2 可以看出，AI 就业冲击接纳、AI 好奇心、知识_应用协同以及 AI 专业关联等特征具有较高的重要性。说明我们在分析 AI 对大学生就业压力的影响时需要重点关注这些因素。比如，AI 专业的关联程度越高，可能意味着大学生所学专业与 AI 的联系更紧密，他们所能感受到的就业压力可能就越大；AI 就业的冲击接纳指出了大学生对 AI 在就业方面影响的认知，若接纳程度越高，则感受到的压力可能越大。但是 AI 创新应用以及 AI 应用举例等特征的重要性相对较低，但依然对模型有一定的贡献，说明这些因素对大学生的 AI 就业压力的感知也有一定的影响。

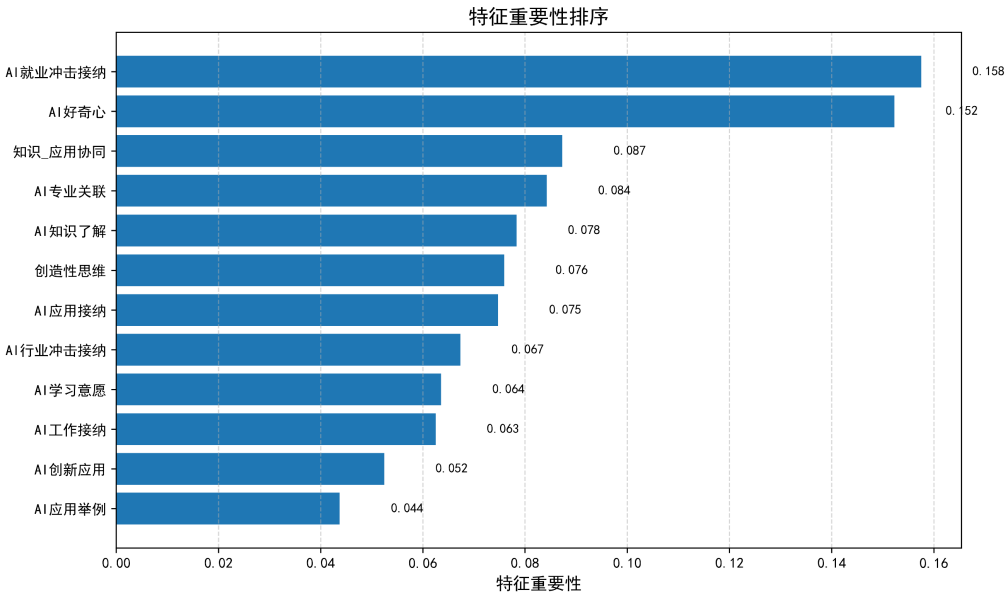


Figure 2. Comparison of feature importance
图 2. 特征重要性对比

通过公式(6)计算得到了 PR 曲线及 AUC-PR 面积值，由图 3 知，PR 曲线下面积 AUC-PR 值为 0.94，接近于 1，说明模型在不同的召回率下整体精确率表现出色。在识别大学生 AI 就业压力高压样本时，可以有效地平衡精确率和召回率，有较强的正样本捕捉能力，能准确区分高压力和低压力样本，在大学生就业压力分类任务中性能良好。如果高校就业指导部门需要关注并且精准定位高就业压力的学生，可

以针对曲线中精确率较高区域对应的阈值；如果希望尽可能找出所有可能面临高压力的学生，可参考召回率较高的部分。这为高校制定针对性就业指导策略、大学生提前规划职业提供了有力支持。

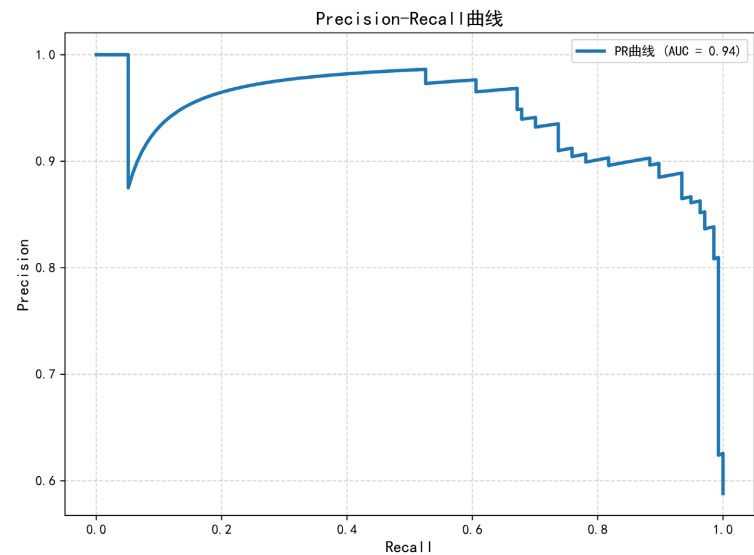


Figure 3. Precision-Recall curves
图 3. Precision-Recall 曲线

通过公式(5)得到 ROC 曲线及 AUC-ROC 面积值，由图 4 知，曲线下面积 AUC-ROC 值为 0.94，接近 1，说明模型在区分大学生 AI 就业压力高、低样本方面能力突出。ROC 曲线的走势较为平稳，无剧烈波动，得出模型在不同样本分布和分类阈值下的表现稳定。这说明该模型在不同的场景下有较好的分类性能，不会因数据的微小变化或阈值的调整而出现大幅波动，有较高的可靠性，因此可作为分析 AI 对大学生就业压力影响的有效工具，为相关研究和就业指导实践提供可靠支持。

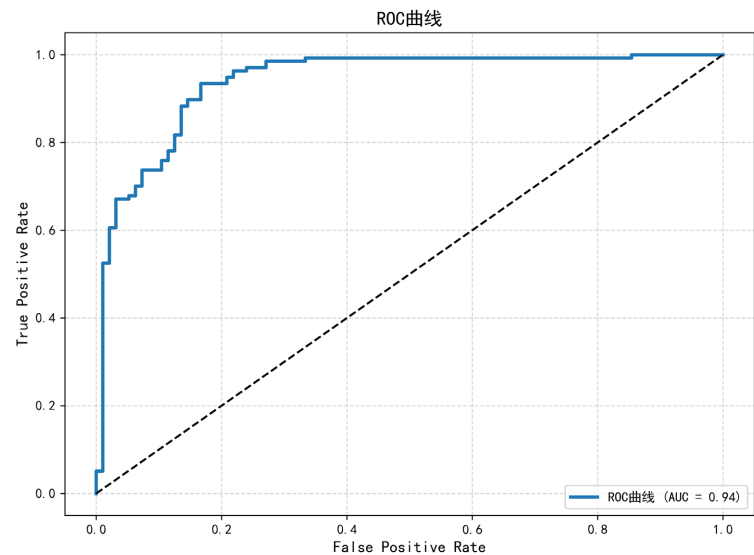


Figure 4. ROC curves
图 4. ROC 曲线

置换重要性也是该模型评价的经典指标之一，可以观察模型是否在该数据集上过拟合。“置换重要性通过破坏特征与标签的关联性来量化特征贡献，是解释随机森林模型的重要工具(Breiman, 2001)。”通过随机打乱某一特征的值，破坏其与标签的关联性，观察模型性能(如准确率、AUC)的下降程度。性能下降越显著，则说明该特征越重要[4]。

Table 4. Permutation importance table
表 4. 置换重要性表

特征	重要性均值	重要性标准差
AI 好奇心	0.065422	0.013090
AI 就业冲击接纳	0.059260	0.012998
AI 专业关联	0.029668	0.006416
AI 知识了解	0.028865	0.007443
AI 学习意愿	0.026711	0.006695
AI 行业冲击接纳	0.025283	0.009249
知识_应用协同	0.021675	0.005724
创造性思维	0.014877	0.005070
AI 应用接纳	0.009862	0.005161
AI 工作接纳	0.009030	0.003841
AI 应用举例	0.005295	0.002766
AI 创新应用	0.003029	0.003233

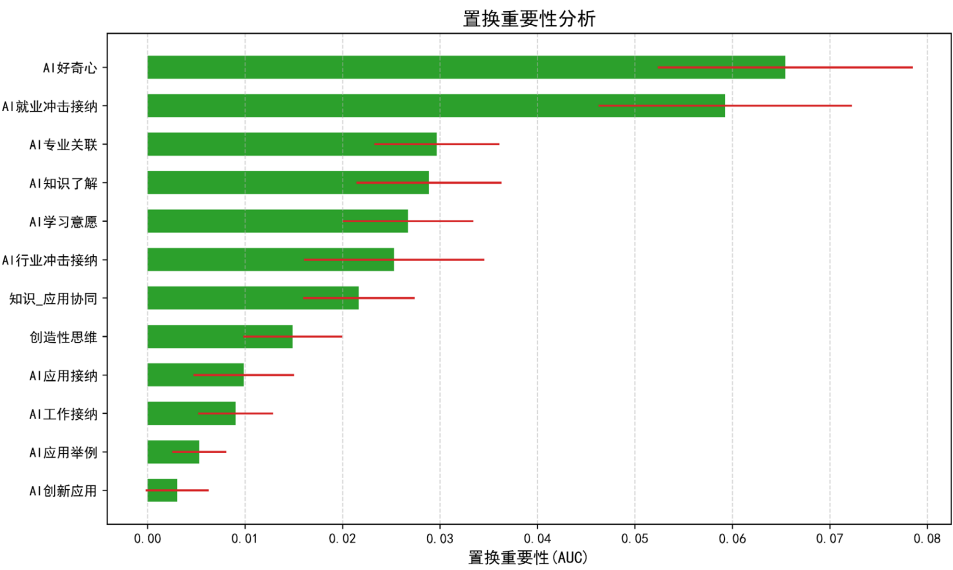


Figure 5. Permutation Importance
图 5. 置换重要性图

对特征值进行置换，重新评估模型准确率，由表 4 中的结果可知；其中，AI 就业冲击接纳、AI 专业关联对数据变化较为敏感，两者的变动分别会导致模型分类准确率下降约 6.422%、5.926%。而 AI 应用

举例、AI 工作接纳只下降约 0.526%、0.303%。由图 5 可知,在此模型中 AI 好奇心的置换重要性最高,表明此特征在区分高低压力组的过程中最为重要。并且通过此图表剔除噪声特征,提升模型泛化能力,使此模型在外部数据集中也能有不错的分类表现。

对分类特征 AI 专业关联、AI 好奇心、AI 知识了解、AI 就业冲击接纳和 AI 学习意愿关于特征取值构筑部分依赖图(简称 PDP 图,展示一个或两个特征对机器学习模型预测结果的边际效应)[5],如图 6 可知,各分类特征随着特征取值的增长,预测概率逐步提高。这说明对 AI 相关的接受程度越高,对高压力组的预测越准确。

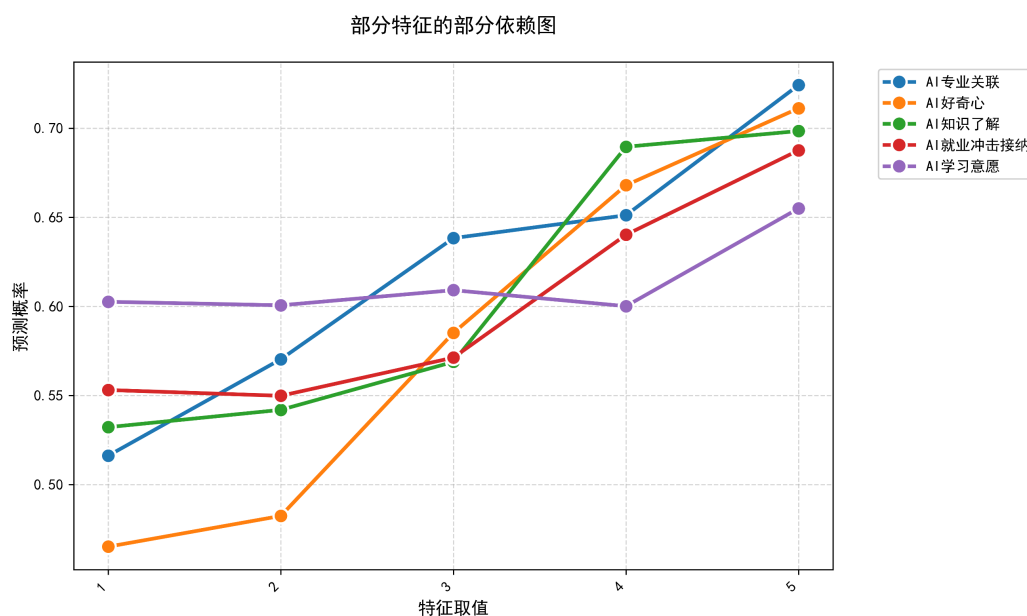


Figure 6. Partial dependency graph
图 6. 部分依赖图

4. 结论

我们通过建立机器学习算法模型分析了 AI 对大学生就业压力的影响,通过模型构建与训练、模型评估等一系列步骤,我们分析出了大学生就业压力的主要影响因素。我们所研究的模型在各项评价指标上表现良好,并且准确率达到 0.8755,精确率达到 0.8857,召回率为 0.9051, F1 的分数为 0.8953, AUC-ROC 值与 AUC-PR 值分别为 0.9399 与 0.9441,证明了模型的有效性以及优越性。随着样本量增大,模型的评价指标可能会有进一步提升[6]。

随着人工智能技术的发展与应用,我国高校毕业生面临前所未有的挑战和机遇,大学生不免为未来感到担忧[7]。通过模型特征重要性图表分析, AI 专业关联、AI 行业冲击接纳、AI 就业冲击接纳等因素对大学生的 AI 就业压力感知影响较大。因此高校可以根据上述指标,加强对于 AI 相关的专业建设以及增加相关课程设置,帮助学生提升专业能力,更好地应对 AI 所带来的就业挑战;同时,高校在开展就业指导工作,需要注重引导学生正确看待 AI 对就业的影响,来提高学生的就业适应能力。对于大学生自身而言,需要关注 AI 技术的发展趋势,并且主动学习相关知识,培养创新思维 and 实践能力,增强自身的就业竞争力。

参考文献

- [1] Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/a:1010933404324>

- [2] 张可, 姚勇, 徐刚, 等. 基于小波变换与随机森林算法的针刺手法脑电分类研究[J]. 生物医学工程研究, 2025, 44(1): 1-7.
- [3] 张可, 李志新, 卢佳琪, 韩菲. 基于深度学习的游客餐饮评论情感分析——以秦皇岛市为例[J]. 数据挖掘, 2024, 14(3): 149-161.
- [4] Strobl, C., Boulesteix, A., Zeileis, A. and Hothorn, T. (2007) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, **8**, Article No. 25. <https://doi.org/10.1186/1471-2105-8-25>
- [5] Friedman, J.H. and Popescu, B.E. (2008) Predictive Learning via Rule Ensembles. *The Annals of Applied Statistics*, **2**, 916-954. <https://doi.org/10.1214/07-aos148>
- [6] Zhou, Z. (2025) Deep Learning and Deep Forest. In: Zhou, Z.-H., Ed., *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall/CRC, 243-269. <https://doi.org/10.1201/9781003587774-11>
- [7] 吕伶俐. 人工智能技术在会展领域的应用对江西高校毕业生就业的影响[J]. 中国会展(中国会议), 2023(24): 106-108.