Computer Science and Application 计算机科学与应用, 2025, 15(1), 46-53
Published Online January 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.151006

Hans汉斯

基于BERT与LightGBM的人岗匹配模型

段雪艳，吕卫东*，冯俊磊，郝月华

兰州交通大学数理学院，甘肃 兰州

收稿日期：2024年12月20日；录用日期：2025年1月16日；发布日期：2025年1月24日

摘　要

在求职招聘市场中，信息不对称导致“逆向选择”，加大了企业招聘和求职者求职的难度。线上招聘平台在疫情时期更加重要，对人岗匹配精度要求更高。传统匹配方式受限，深度学习技术特别是BERT模型和集成模型受到关注。当前学者在研究人岗匹配问题时，采用常见的TF-IDF词向量表示方法和Word2Vec词向量表示方法来对中文文本进行表征，但是由于科学的进步，当下用BERT模型能更好地读取文本语义，因此本文将BERT模型引入到人岗匹配领域中，采取了基于BERT模型的词向量表示和LightGBM模型的人岗匹配方法，以提升匹配精度和效率，与多种机器学习模型的预测结果相比较之后，最终发现，在这两种方法的结合下，在本文所构建的人才是否投递模型中的精确度达到了0.886，在岗位是否认可模型中的精确度达到了0.926，由这两个模型的效果可以看出BERT模型和LightGBM模型的结合，可以为招聘平台提供精准模型。

关键词

BERT模型，人岗匹配，LightGBM模型

A Person-Job Matching Model Based on BERT and LightGBM

Xueyan Duan, Weidong Lyu*, Junlei Feng, Yuehua Hao

School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou Gansu

Received: Dec. 20th, 2024; accepted: Jan. 16th, 2025; published: Jan. 24th, 2025

Abstract

In the job recruitment market, information asymmetry leads to “adverse selection”, which increases the difficulty for both enterprises in hiring and job seekers in finding employment. Online recruitment platforms have become even more crucial during the pandemic, placing higher demands on

*通讯作者。

文章引用: 段雪艳, 吕卫东, 冯俊磊, 郝月华. 基于 BERT 与 LightGBM 的人岗匹配模型[J]. 计算机科学与应用, 2025, 15(1): 46-53. DOI: 10.12677/csa.2025.151006

the accuracy of person-job matching. Traditional matching methods are limited, and deep learning technologies, especially the BERT model and ensemble models, have garnered attention. In current research on person-job fit, scholars often represent Chinese text data using common methods such as TF-IDF word vectors and Word2Vec word vectors. However, due to advancements in science and technology, the BERT model is now better at capturing textual semantics. Therefore, this paper introduces the BERT model into the field of person-job fit. This paper proposes a person-job matching method based on the BERT and ensemble models to improve matching accuracy and efficiency. After comparing the prediction results with various machine learning models, it was ultimately found that with the combination of these two methods, the accuracy of the talent submission model constructed in this paper reached 0.886, and the accuracy of the job acceptance model reached 0.926. The effectiveness of these two models demonstrates that the combination of the BERT model and the LightGBM model can provide a precise model for recruitment platforms.

Keywords

BERT Model, Person-Job Matching, LightGBM Model

1. 引言

在当今信息泛滥的时代，招聘市场广阔无垠，企业如同航船。在这片市场中，找到合适的人才与工作岗位绝非易事，特别是在激烈的竞争环境中。传统的人岗匹配方法难以精确捕捉文字背后的深层联系。随着科技的进步，自然语言处理(NLP)与机器学习技术，为我们开辟了新路径。BERT 模型凭借出色的文本理解能力，能深入文本，捕捉关键语义信息。这为重新审视人岗匹配带来了新视角和有力工具。

然而，单一的 BERT 模型虽强，却也有其局限性。正如任何单一工具都有其适用范围一样，BERT 模型在面对复杂多变的人岗匹配任务时，也可能遭遇瓶颈。因此，我们想到了集成模型这一策略，它能够与不同模型协同作业，通过综合各模型的信息与优势，制定出更加精准、高效的模型。正是基于这样的思考，提出了基于 BERT 模型和各类机器学习模型相结合的人岗匹配方法。这一方法旨在将 BERT 模型强大的文本理解能力与集成模型的综合优势相结合，构建出一个更加智能、高效的人岗匹配系统。在接下来的内容中，将深入探讨这一方法的理论基础、实现细节以及实际应用效果，以期为人岗匹配领域的研究和实践贡献一份力量。

2. 理论知识

BERT 是 Google 于 2018 年提出的基于 Transfommers 的双向编码器结构模型[1]。它是一种深度学习语言模型，能够高效地学习通用语言表征。其核心在于 Transformer 编码器，结构如图 1(a)所示[2]。

在处理输入文本的过程中，首先会进行编码处理并融入位置信息，进而构建出输入向量。这些向量接着被传递给 Transformer Encoder 进行处理，其中每个编码器默认包含六个重复的编码器层。每个编码器层则由一个多头自注意力机制模块和一个全连接的前馈网络模块构成，模块间采用残差连接与层标准化技术以增强训练效果。经过这一系列处理，最终生成与输入向量维度相匹配的输出向量，这一过程实质上是对输入文本中每个字符的语义向量进行转换，使其增强为等长且具有更丰富语义信息的向量。

如图 1(b)所展示，BERT 模型是通过多层 Transformer Encoder 的堆叠构建的。它接收的是文本中每个字或词的初始词向量作为输入，而输出的是这些字或词在综合了整篇文本语义后的向量表达。BERT 模型分为两个不同规模的版本：BERT (base)和 BERT (large)。其中，BERT (base)版本配置了 12 个自注意力头以及 12 层的 Transformer Encoder，总共包含大约 1.1 亿个参数；而 BERT (large)版本则更为庞大，它配备了 16 个自注意力头和 24 层的 Transformer Encoder，参数总数高达 3.4 亿。

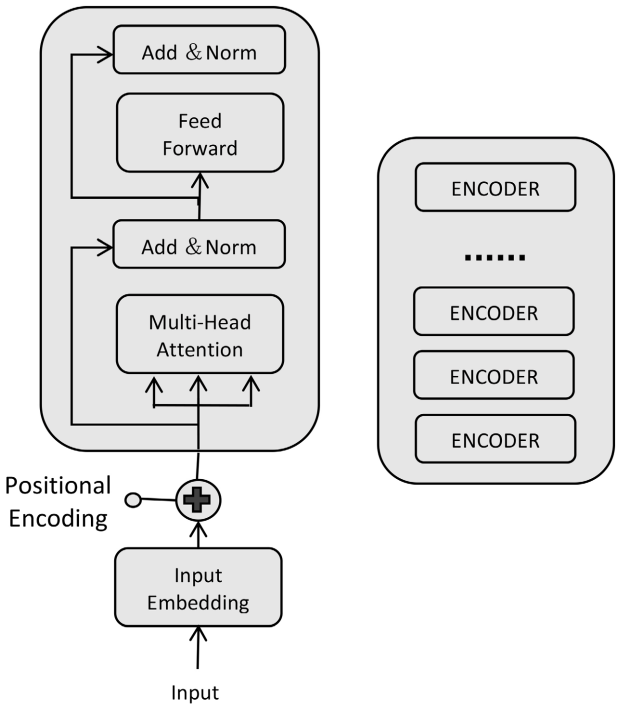

Figure 1. Classification model based on BERT pre-trained model
图 1. BERT 预训练模型的分类模型

鉴于 BERT 模型参数规模庞大，达到亿级级别，其训练过程不仅需要海量的数据支持，还对硬件资源提出了极高的要求，这对于大多数普通研究者而言构成了难以逾越的障碍。尽管像 Google 这样的行业巨头已经公开了他们的预训练成果，但现有的模型维护与应用体系尚不完善，用户在实际使用时往往会遇到诸多挑战。为了改善预训练语言模型的使用体验，Huggingface 公司创建了一个大型预训练模型社区，该社区汇聚了多种预训练模型及其相关数据资源。通过提供高效且易于使用的统一接口，该社区显著降低了模型的使用难度。此外，社区还整合了基准测试数据集，使用户能够方便地对比不同模型的性能，从而更加轻松地选择和使用适合自己的模型。

人岗匹配模型分为两部分：前端是从 Hugging Face 上挑选的 BERT 预训练模型，后端是一个分类器[2]，见图 2。

前端选用了名为 chinese RoBERTa L-12H-768 的预训练模型[3]，其训练素材源自 Common Crawl 的中文数据集，涵盖了大约 100GB 的高质量中文预训练文本。此模型的 BertModel 层架构包含三个核心子层：BertEmbeddings (嵌入层)、BertEncoder (编码层)以及 BertPooler (池化层)。

BertEmbeddings 具有 512*768 的位置嵌入维度，使得模型能处理最长为 512 个字符的文本。BertEncoder

由 12 层 BertLayer 堆叠而成，每层特征维度均为 768，且配备了 12 个自注意力头。完成处理后，[CLS]标记的输出向量会传入 BertPooler 层进行深化加工，以精炼出输入文本的核心特征。

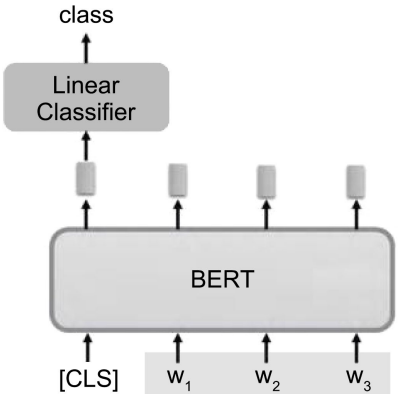

Figure 2. Schematic diagram of the BERT model
图 2. BERT 模型示意图

3. 实验分析

3.1. 数据来源

本文所采纳的数据源自天池竞赛平台所主办的“大数据智能云编程大赛——人岗智能匹配挑战赛”。该赛事中所应用的职位信息与简历资料均基于真实世界的数据集构建。此数据集全面包含了求职者的个人资料、岗位的具体需求信息，以及求职者与岗位之间发生的互动行为记录(图 3)。

数据类型	特征
求职者	用户 id、现居住地、期望工作城市、期望行业、期望职类、期望薪水、最近工作行业、最近工作职类、最近新水、学历、年龄、开始工作年、工作经历
岗位	岗位 id、职位标题、公司名称、城市、职位子类、需求人数、最高月薪、最低月薪招聘开始日期、招聘结束日期、是否出差、工作年限要求、最低学历、最高学历、是否要求管理经验、语言需求、岗位描述
行为数据	用户 id、职位 id、求职者是否浏览岗位、求职者是否投递简历、招聘者是否认可求职者

Figure 3. Dataset information
图 3. 数据集信息

3.2. 数据预处理

原始数据集广泛收集了 4500 位求职者的信息 269,534 条岗位记录以及 700,938 项行为记录。预处理流程涵盖了数据清洗(包括处理缺失值和重复值)、数据合并、数据规范化(即将文本信息转化为数值格式)，以及针对长文本字段的专门处理，具体包括中文词汇的分割、停用词的去除以及中文文本相似性的评估等，这一系列步骤旨在提升样本数据的标准化水平。同时最大限度地保留原始数据的完整性。这些措施为后续的特征提取、数据挖掘及模型训练奠定了坚实基础，旨在从数据维度优化算法性能，确保达成预

期成效。预处理完成后，数据集缩减为包含 4122 名求职者、33,340 个岗位及 53,458 条行为记录。

3.3. 特征提取

样本数据中蕴含丰富的非直接显示信息，这些信息往往需要通过特征提取过程来发掘，以便从样本中提炼出更多有价值的内容，并将其转换成算法能够识别的特征形式。特征提取不仅能够有效地实现数据集的维度缩减，还能简化模型训练过程、增强模型精确度，并削弱噪声数据干扰。

首先，从岗位数据和求职者数据两个方面着手进行基础特征的抽取。在经过规范化的岗位样本数据集中，主要呈现了与岗位招聘相关的各项要求，涵盖了 11 个特征维度。而在求职者简历样本数据集中，则主要展示了求职者的基本信息及其对岗位的期望，共包含了 15 个特征维度。此外，关于求职者与岗位招聘方之间的交互行为信息，还提取了 4 个特征维度。

提取完基础特征再对文本特征进行提取，对于数据集中的文本数据“岗位要求”和“工作经历”两部分，在本研究的实验中，用 BERT 预训练模型来捕获词语的上下文动态嵌入表示，并将模型中的[CLS]标记所输出的向量作为人岗匹配模型的特征向量。这里的[CLS]标记能够概括并输出包含全局信息的文本表征[3]。

通过分析人岗信息数据集中的基本特征，可以观察到招聘岗位信息与简历信息之间存在的相似的信息字段较多。针对这些共有的信息，进行专门的特征提取操作，旨在构建出交叉特征。这一步骤不仅有助于更有效地降低特征维度，还能促进特征群体的系统化构建。也就是通过整合岗位与人才数据中的共同信息，提炼出关键的交叉特征，从而实现数据的精简与特征的高效利用。运用上述方法，对整合后的样本数据集进行了处理，针对其中的数值型与字符型变量，分别实施了特征提取与转换操作，最终成功引入了以下 6 个新的交叉特征(图 4)。

交叉特征名称	交叉特征说明
In_work_year_min	是否满足最低工作经验
In_work_year_max	是否满足最高工作经验
In_desire_salary_min	是否达到最低期望薪资
In_desire_salary_max	是否达到最高期望薪资
degree	是否满足最低学历
in_desire_city	期望城市是否吻合

Figure 4. Cross features
图 4. 交叉特征

在原始数据集中，每位求职者都与多个岗位存在关联，相应地，每个岗位也与一组求职者相匹配，从而构成了大量的人才－岗位配对实例。因此，可以从人才和岗位两个维度出发，依据用户 ID (user ID)和岗位 ID (position ID)对各类变量实施分组统计。具体做法包括计算最大值、最小值、平均值以及标准差等统计量[4]，以此来评估：从岗位视角看，人才的相对竞争力；以及从人才视角审视，岗位的吸引力或竞争力。这样的分析有助于深入理解人才与岗位间的竞争关系。

经过上述特征提取流程，我们从样本数据集中整理出了 235,474 条数据，涵盖了 11 个岗位基础特征、15 个人才简历基础特征、4 个行为特征、20 个文本特征，以及 6 个由岗位与人才信息融合而成的交叉特

征。此外，还计算了 42 个统计特征。为了精简特征集并避免信息冗余，对交叉特征进行了整合处理，最终构建了一个包含 92 个独特特征的特征群，以供后续分析使用。

4. 模型构建与结果分析

经过前期的特征提取步骤，将行为特征中的“delivered”标记(0 代表未投递，1 代表已投递)设定为判断人才是否进行模型投递的目标变量[4]，同时，将“satisfied”标记(0 表示不认可，1 表示认可)设定为评估岗位招聘者是否满意模型的目标变量[5]。随后，采用基础特征、文本特征、人才与岗位的交互特征，以及基于人才视角的岗位统计特征，作为构建预测模型的关键输入变量[5]。在此基础上，我们将人才投递行为的预测与岗位招聘者满意度的预测均转化为二分类任务，以便于模型的训练与后续的预测评估工作。

在构建二分类模型时，多种成熟技术可选，如统计学的判别分析、逻辑回归(LR)，以及机器学习的 SVM、NN 和决策树等。鉴于变量多且定性变量占比大，这些技术均可考虑。决策树家族的方法在此情境下显得尤为适用。实证分析已表明决策树方法具有优势，故本文选用决策树家族模型，包括决策树、随机森林、XGBoost 和 LightGBM 等[6]。

利用所选样本数据集，按 7:3 的比例划分为训练集和测试集。针对人才投递和招聘者认可这两个角度展开二元分类任务，我们训练了一系列二分类模型，得出了以下训练结果(表 1、表 2)：

Table 1. Prediction results of various models for whether talents have been delivered
表 1. 针对人才是否投递(delivered)各模型预测结果

预测模型	Accuracy	Precision	Recall	F1	AUC
决策树	0.771	0.339	0.541	0.405	0.709
随机森林	0.842	0.722	0.829	0.771	0.839
XGBoost	0.870	0.814	0.834	0.823	0.836
LightGBM	**0.886**	**0.776**	**0.750**	**0.760**	**0.848**

Table 2. Prediction results of various models for whether recruiters are satisfied
表 2. 针对招聘者是否认可(satisfied)各模型预测结果

预测模型	Accuracy	Precision	Recall	F1	AUC
决策树	0.877	0.766	0.740	0.750	0.844
随机森林	0.899	0.737	0.842	0.785	0.896
XGBoost	0.925	0.812	0.866	0.836	0.908
LightGBM	**0.926**	**0.821**	**0.867**	**0.840**	**0.919**

根据上述模型训练的结果，采用了 ROC 曲线和 AUC 值作为评价指标，来评估各模型的预测性能[6]。其中，AUC 值能够反映模型在不同分类阈值下的性能表现。综合考虑决策树(DT)、随机森林(RF)、XGBoost 和 LightGBM 这四种算法模型，分别针对两个不同目标变量的预测任务，最终绘制出了各类模型的 ROC 曲线，具体图示如下图 5、图 6。

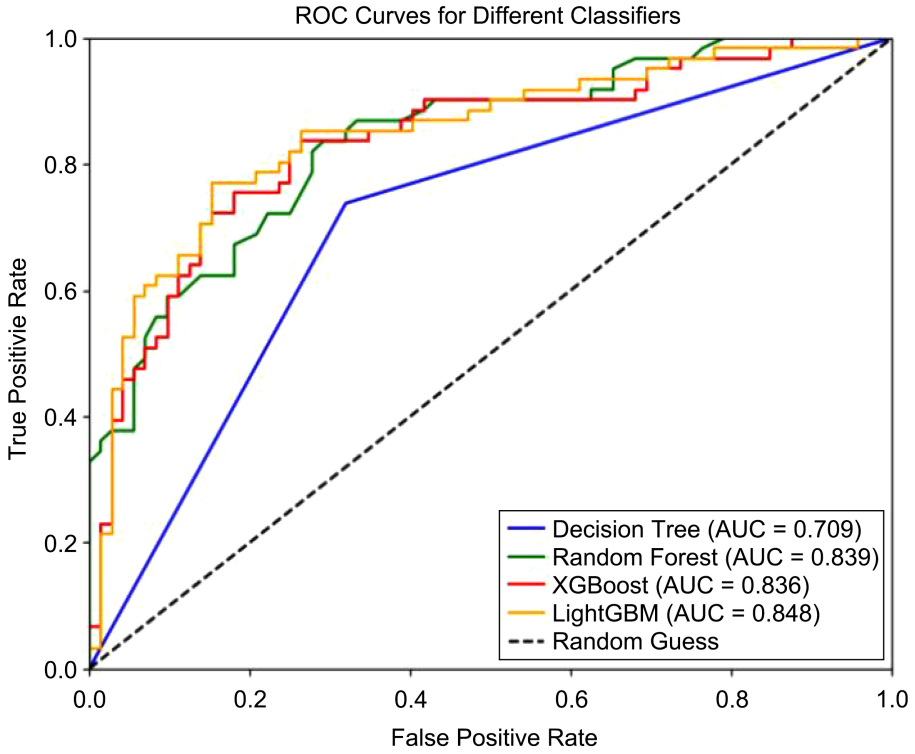

Figure 5. ROC curve of the prediction model for whether talents have been delivered
图 5. 人才是否投递预测模型的 ROC 曲线

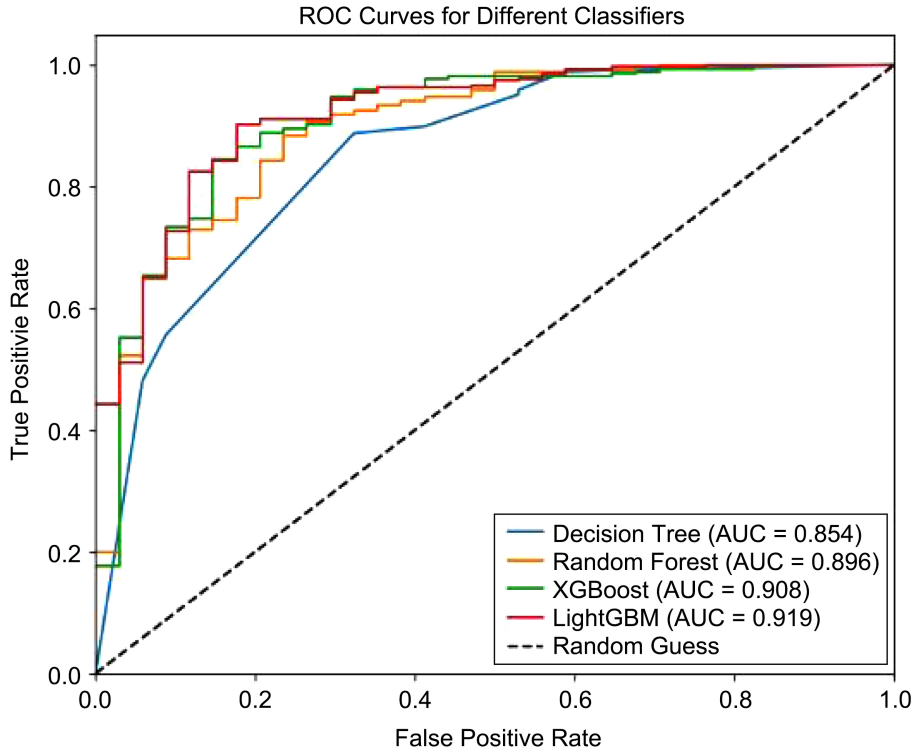

Figure 6. ROC curve of the prediction model for whether the job recruiter is satisfied
图 6. 岗位招聘方是否认可预测模型的 ROC 曲线

将这两类预测模型对应的 AUC 值整理如表 3 所示。

Table 3. AUC values of various models
表 3. 各模型 AUC 值

模型名称	Auc (delivered)	Auc (satisfied)
决策树(DT)	0.709	0.854
随机森林(RF)	0.839	0.896
XGBoost	0.836	0.908
LightGBM	**0.848**	**0.919**

观察表 3 可知，在预测人才是否投递(Delivered)的模型中，除决策树外，其他三个模型的评估指标相近，但 LightGBM 模型的训练效果略胜一筹，优于随机森林和 XGBoost。而在预测岗位招聘者是否认可(Satisfied)的模型中，LightGBM 模型的表现同样更为出色，其训练效果明显优于其他三个模型，展现出最佳的性能。因此，我们选择了 LightGBM 模型作为最终的人岗匹配模型。

5. 结论

LightGBM 算法包含众多参数，这些参数的选择以及结合外部交叉验证进行模型调优，会显著影响模型的预测效果和性能[7]，也规避了模型数据过拟合现象，并且有效地融合了语义相似度与 BERT 模型[8]，这才使得基于 BERT 模型与 LightGBM 模型的人岗匹配取得了较为理想的效果，从而实现人岗匹配智能推荐。然而，由于本文所收集的数据仅限于人才、岗位及行为三方面的历史记录，缺乏相应的验证行为数据，因此模型融合的实际效果还需在实践中深入验证。

参考文献

[1] 李冬梅, 朱朝阳, 李丽, 等. 基于 BERT 实现基础医学专业术语智能提取系统[J]. 基础医学教育, 2024(11): 1002-1007.
[2] 查佳凌, 金薇, 徐呈宙, 等. 基于 BERT 的青少年心理健康预警模型[J]. 中国数字医学, 2024, 19(10): 101-106.
[3] 张红. 语义相似度与 BERT 模型融合的多标签文本自适应分类方法[J]. 微型电脑应用, 2024, 40(3): 49-52.
[4] 马驰. 人才招聘平台的人岗匹配模型优化设计及应用研究[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2021.
[5] 刘敏. 融合双方偏好信息的人职匹配研究[D]: [硕士学位论文]. 太原: 山西大学, 2023.
[6] 段丹丹, 唐加山, 温勇, 等. 基于 BERT 模型的中文短文本分类算法[J]. 计算机工程, 2021, 47(1): 79-86.
[7] 左玉倩. 基于 BiLSTM 和 XGBoost 的人岗匹配方法研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2022.
[8] 吴越, 孙海春. 基于图神经网络的知识图谱补全研究综述[J]. 数据分析与知识发现, 2024, 8(3): 10-28.