参赛队号：（由大赛组委会办公室填写）

# 2021年（第七届）全国大学生统计建模大赛

| | |
|---|---|
| 参赛学校： | 江西财经大学 |
| 论文题目： | 基于 LightGBM 和 BP 神经网络的<br>互联网招聘需求分析与预测 |
| 参赛队员： | 黄婷、叶妍言、刘倩茵 |
| 指导老师： | 余达锦、刘庆 |

# 基于 LightGBM 和 BP 神经网络的
# 互联网招聘需求分析与预测

## 目录

# 表目录

# 图目录

# 摘要

就业是民生之本，是发展之基，也是安国之策。2020 年新冠肺炎疫情的爆发，稳就业成为应对疫情、稳定社会的重要保障之一。随着数据新动能的发展，互联网招聘为招聘者和应聘者提供不限于时空的全局视角，因此本文从该角度出发对招聘者和应聘者需求进行统计分析预测，以期缓解就业难、招聘难的困境。

本文基于近年来各在线招聘网站所发布的招聘数据并结合数据新动能下转型升级的三个金融行业、互联网行业、生产制造行业，采用 Pearson 相关系数检验初步筛选后运用灰色关联分析进一步进行指标筛选，最后对企业招聘中招聘者关注的浏览量运用 LightGBM 模型进行浏览量特征重要性分析，对就业形势中应聘者关注的薪资运用 BP 神经网络预测模型对于薪资进行预测，并进行模型精度对比，得出数据新动能下三个行业的薪资统计分析预测。

经研究得出关于企业招聘浏览量，金融行业薪资水平，互联网行业薪资水平，生产制造行业薪资水平的影响因素及重要程度。基于以上分析结论，本文在互联网招聘市场中对招聘者与应聘者需求提出以下对策建议：第一，对于企业，招聘者应根据岗位浏览量合理设置招聘要求；第二，对于金融行业，应聘者应根据学历因素合理考虑就业地域；第三，对于互联网行业，应聘者应根据学历因素合理考虑公司性质；第四，对于生产制造行业，应聘者应根据公司所在地合理考虑公司性质。



关键词：数据新动能；互联网招聘；就业形势；LightGBM 模型；BP 神经网络

# 一、绪言

## （一）研究背景与意义

习近平总书记指出：“就业是最大的民生，是最大的民生工程、民心工程、根基工程。”中共中央、国务院印发的《关于构建更加完善的要素市场化配置体制机制的意见》明确要破除阻碍要素自由流动的体制机制障碍，提出了土地、劳动力、资本、技术、数据五大要素领域的改革方向。伴随着我国数据新动能的高速发展，就业的重要意义不容忽视。

2020 年我国农民工达 2.9 亿、高校毕业生达 874 万，加之新冠肺炎疫情爆发以来，经济下行幅度大，企业被迫减少人力资源支出，招聘需求减少，社会失业人口增加，稳就业如何发力，成为时代热点。习近平总书记高度关注疫情对就业形势的影响，就统筹推进疫情防控和经济社会发展工作作出一系列重要指示，为做好疫情防控中的就业工作提供了科学指引。

互联网招聘为招聘者和应聘者提供不限于时空的全局视角，招聘者可以在互联网平台上发布招聘信息，而应聘者在互联网招聘平台上也能依据自身需求找到自己满意的岗位。但目前互联网招聘也存在招聘者和应聘者需求不对称问题，基于此本文从该视角出发，对于互联网招聘市场应聘者和招聘者需求进行分析与预测，以期解决需求不对称导致的就业难、招聘难问题。

## （二）研究现状

互联网招聘的概念是在 20 世纪 80 年代中期首次引入劳动力管理市场的，在 20 世纪 90 年代随着改革开放进入中国。而数据新动能发展时期背景下，劳动力作为最活跃的生产要素，就业是各级政府宏观调控的重要目标，互联网大数据以其更快、更细、更敏锐的优点帮助提高宏观决策的灵敏性和时效性。互联网求职与招聘是求职者通过网络招聘平台与招聘方进行条件匹配的过程，下面从求职者、用人单位和网络招聘平台三方对互联网招聘进行综述与分析。

1.互联网招聘对求职者的相关研究

近年来，解决就业问题是我国的一项重大工作。传统招聘由于招聘渠道单一、招聘成本高昂、无法更为精确地进行应聘者能力考察等问题在招聘活动中逐渐乏力（徐汝婷,蔡晓晶，2015）。互联网以其开放性、平等性、互动性、智能性对劳动力需求产生了巨大的影响，引发了产业结构升级、商业模式创新和创业机会增加（黄敬宝，2015）。研究表明，在互联网寻找工作可以使失业期缩短 25%。互联网招聘以其高可信度、不受地域限制、低成本等优势一跃成为新一代的“宠儿”（李浇和倪燕茹，2017）。互联网招聘通过强化信息沟通，大幅提升了求职者与招聘方的匹配率，求职者能最大程度上找到最合适的岗位，招聘方也可以选择到最合适的人才。2015 年，网络招聘已成为我国的主要招聘方式。数据新动能发展时期，对人才也有了更高的要求，拥有互联网思维、业务技术能力较强、网络操作能力较强、资源整合能力较强的优秀人才会具有更强的就业能力（黄敬宝，2015）。然而，在目前的劳动力要素市场上充斥着大量虚假信息，而国家几乎没有针对网络招聘的政策法规和管理办法，求职者可能是这些虚假信息的发布者，如求职简历“注水”（刘谢彪，2020），同时也可能是虚假信息的受害者，即使权利被不法分子欺骗也无法得到保障（尹淑华，2021），给招聘方、求职者、就业市场秩序和整个社会带来了危害。

2.互联网招聘对用人单位的相关研究

吴文艳（2014）提出，招聘管理是“组织基于生存和发展的需要，根据组织人力资源规划和工作分析的数量与质量要求，采用一定的方法吸纳或寻找具备任职资格和条件的求职者，并采取科学有效的选拔方法，筛选出符合本组织所需合格人才并予以聘用的管理活动。”招聘管理补充企业的人力资源，不仅对企业后续的各项活动有着直接的影响，同时也是树立企业形象和扩大企业知名度的过程，具有非常重要的地位。互联网技术对企业招聘管理也带来许多便利：（1）招聘信

息发布更快，成本更低。（2）简历筛选精准方便。（3）测评方式灵活（李燕萍，齐伶圆，2016）。企业招聘的目标是吸引有天赋的人才，并将他们留在企业中。为了实现这一目标需要依照流程进行电子招募（E-recruiting）、电子甄选（E-selecting）和电子评估（E-evaluation）。在大数据时代，大数据技术在人力资源招聘活动中的应用已成为必然趋势，由此又产生了基于机器学习、大数据应用技术对互联网招聘效率提升研究。企业可以通过构建人才画像、岗位画像提升人才选拔的精确度（韩琰，2020），与机器学习算法相结合自动筛选简历（尹源，2020），提高人才招聘效率。

3.互联网招聘对网络招聘平台的相关研究

招聘网站为企业和求职者建设起双方信息匹配的中介平台。对企业的服务，包括招聘广告和招聘信息发布、简历查询等；对个人的服务，则是提供职位搜索、简历投递等。招聘网站主要包括综合性招聘网站、专业性招聘网站、区域性招聘网站（谷彬，2016）。在中国，信息覆盖较为全面的综合性招聘网站中最具有代表性的是智联招聘、前程无忧、中华英才，这三家公司也是最早开始投身中国网络招聘的企业（侯隽，2015）。专业性招聘网站包括行业招聘网站和职业招聘网站两个角度，前者反映生产供给角度的分工细化与专业化，后者反映了市场需求较为突出的热点领域。区域角度来看，则呈现五大特征：精确聚焦区域、范围交叉融合、聚焦地区主导行业、体现地区特色、与中国文化相结合。面对企业和招聘者需要的两端个性化、精准化、高效化的要求，招聘网站围绕着招聘做一系列的服务:人力资源服务、求职面试指导、简历服务、职业发展规划服务等（卞文志，2018），这些边缘化服务在无形中极大地提高了用户黏性。大数据智能分析系统更是能在瞬间高效完成数据收集、分析到可视化的过程，极大地提高了网络招聘平台的信息搜集处理能力，为用户呈现高度精准匹配的数据信息（罗宜春，2020）。然而，在部分招聘平台急功近利的影响下,一些招聘网站存在隐私泄漏、倒卖用户

简历的问题（张向阳，2021）。

综上所述，国内外学者对互联网招聘市场已取得显著进展，形成了丰硕成果，这为本论文提供了重要的理论参考。基于数据新动能下互联网招聘情况，还存在以下进一步探索的空间：（1）对于应聘者，如何帮助其正确定位自身条件、有效寻求更合适的职位。（2）对于招聘者，如何选拔到更切合企业发展的应聘者、同时降低成本提高效率。

### （三）研究思路

基于当代数字经济大环境背景，面对当前互联网市场应聘者和招聘者需求不对称的现状，本文运用近年来各在线招聘网站所发布的招聘数据并结合数据新动能下转型升级的三个金融行业、互联网行业、生产制造行业，采用 Pearson 相关系数分析初步筛选后运用灰色关联分析进一步进行维度筛选，最后对企业招聘中招聘者关注的浏览量运用 LightGBM 模型进行特征重要性分析，对就业形势中应聘者关注的薪资运用 BP 神经网络预测模型对于薪资进行预测，并进行模型检验与修正，得出新动能下三个行业的薪资和浏览量的分析与预测。

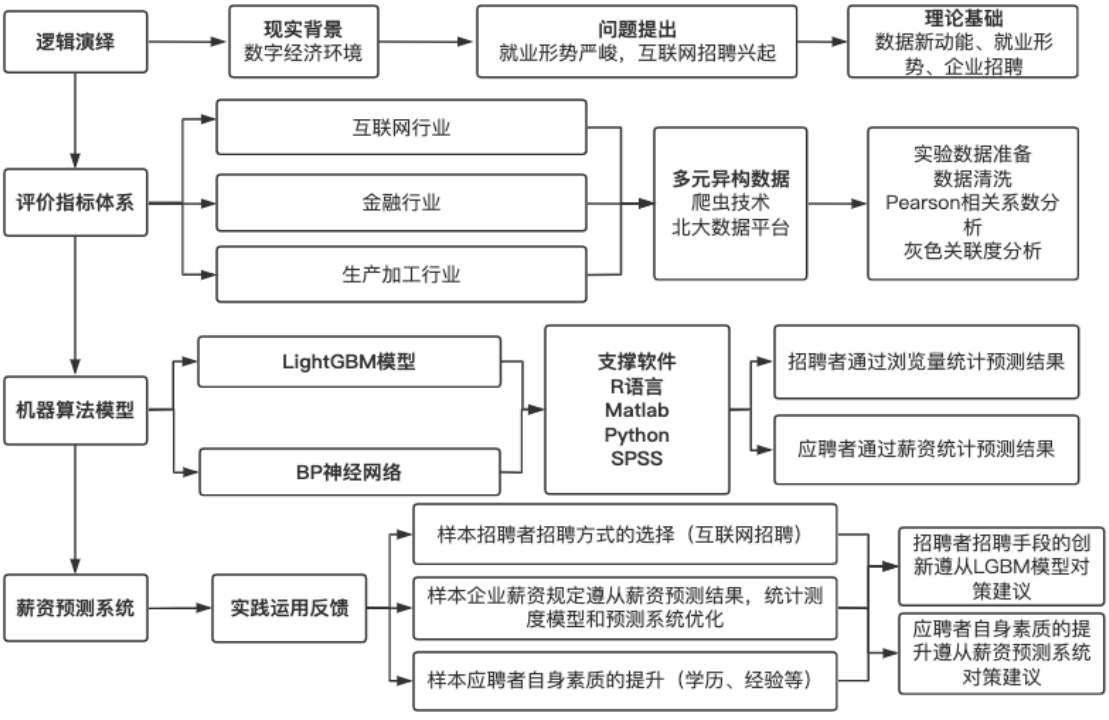



图 1　研究技术路线图

## 二、指标的选取与数据的处理

### （一）数据来源

本文数据通过对某数据平台的数据进行爬取，总共得到 1007894 条数据。数据预处理以 excel 为主，Python、R 为辅，完成原始数据去重区空以及数值转换等数据预处理工作之后进行分层随机抽样得到剩下 40000 条数据进行统计分析。

对于异常值的处理，学历、职位、行业等因素使用删除异常值方法处理，经验年数、工资上下限因素使用计算平均值方法进行处理。分层抽样法，也叫类型抽样法。将总体单位按其属性特征分成若干类型或层，然后在类型或层中随机抽取样本单位。分层抽样法的特点是通过划类分层，增大了各类分层抽样中单位间的共同性，容易抽出具有代表性的调查样本。该方法适用于总体情况复杂，各单位之间差异较大，单位较多的情况。分层随机抽样的程序是把总体各单位分成两个或两个以上的相互独立且各具特点的完全的组，再从两个或两个以上的组中分别进行随机抽样。分组的标志或特点与所关心的总体特征相关。“所学非所用”不利于充分发挥人力资本的潜在价值（郭睿，2019），本文以学历作为属性特征进行分层，将不同学历分出不同层，按各学历占总数据的比例在每一层中随机抽样，得出 40000 条数据。

并通过划分行业来分别选取每个行业中的指标进行分析预测，金融行业的发展是一个国家经济发展的重要支撑（高景文，2019），互联网行业则为数字化时代背景下一个重要的行业支撑（周蕴慧，2021），生产制造行业的转型升级也是当今时代面临的重大课题（江小涓，2020），这三个行业都对数据新动能背景下招聘与就业需求不对称的统计分析研究具有一定意义，因此本文选取这三种行业进行统计分析预测。

而对于大多数互联网应聘者而言，薪资是众多被考虑因素中的重中之重，是其劳动回报的直接体现（Kristin L ，2018），对于企业而言，应聘者的薪资与其

经营的利润以及成本是直接相关的关系。因此选取三个行业薪资平均值与其他指标进行分析。

### （二）指标选取

1. Pearson 相关系数检验

Pearson 相关系数是用协方差除以两个变量的标准差得到的，虽然协方差能反映两个随机变量的相关程度（协方差大于 0 的时候表示两者正相关，小于 0 的时候表示两者负相关），但是协方差值的大小并不能很好地度量两个随机变量的关联程度，对于标准化后的数据求欧氏距离平方并经过简单的线性变化，也就是 Pearson 系数，我们一般用欧式距离来衡量向量的相似度，但欧式距离无法考虑不同变量间取值的差异。加之，Pearson 相关系数适用于高维度检验，而未经升级的欧式距离以及 cosine 相似度，对变量的取值范围是敏感的，在使用前需要进行适当的处理。因此在对变量间进行相关性检验时，本文优先采用 Pearson 相关系数检验去研究经验，学历，公司所在地，公司性质，职位分别和薪资平均值之间的相关关系，使用 Pearson 相关系数去表示相关关系的强弱情况。具体分析可知：

①金融行业：经验、学历、职位、公司所在地呈现显著性

表 1　金融行业 Pearson 相关系数

| | 经验 | 学历 | 职位 | 公司所在地 | 公司性质 |
|---|---|---|---|---|---|
| 薪资平均值 | 0.055** | 0.114** | 0.064** | -0.132** | -0.028 |
| * p<0.05 ** p<0.01 | | | | | |

②互联网行业：经验、学历、职位、公司所在地、公司性质呈现显著性

表 2　互联网行业 Pearson 相关系数

| | 经验 | 学历 | 公司所在地 | 公司性质 | 职位 |
|---|---|---|---|---|---|
| 薪资平均值 | 0.027** | 0.126** | -0.160** | -0.123** | -0.026* |
| * p<0.05 ** p<0.01 | | | | | |

③生产制造行业：经验、学历、公司所在地、公司性质呈现显著性

表 3　生产制造行业 Pearson 相关系数

| | 经验 | 学历 | 公司所在地 | 职位 | 公司性质 |
|---|---|---|---|---|---|
| 薪资平均值 | 0.104** | 0.081** | -0.082** | 0.013 | -0.098** |

* p<0.05 ** p<0.01

2. 灰色关联分析

基于 Pearson 相关系数检验得出的结果，本文进一步对具有显著性的各个特征值进行选取。运用灰色关联分析对于研究指标进行进一步选取，研究各因素对薪资的影响大小关系，得出结果如下：

①金融行业：公司所在地、职位

表 4　金融行业灰色关联度结果

| 评价项 | 关联度 | 排名 |
|---|---|---|
| 经验 | 0.640 | 4 |
| 学历 | 0.644 | 3 |
| 职位 | 0.670 | 2 |
| 公司所在地 | 0.989 | 1 |

从上表可以看出：针对本次 4 个评价项，公司所在地的综合评价最高（关联度为：0.989），其次是职位（关联度为：0.670）。

②互联网行业：学历、公司性质

表 5　互联网行业灰色关联度结果

| 评价项 | 关联度 | 排名 |
|---|---|---|
| 经验 | 0.902 | 5 |
| 学历 | 0.928 | 1 |
| 职位 | 0.909 | 3 |
| 公司所在地 | 0.907 | 4 |
| 公司性质 | 0.927 | 2 |

从上表可以看出：针对本次 5 个评价项，学历的综合评价最高（关联度为：0.928），其次是公司性质（关联度为：0.909）。

③生产制造行业：公司所在地、公司性质

表 6　生产制造业关联度结果

| 评价项 | 关联度 | 排名 |
| --- | --- | --- |
| 经验 | 0.950 | 4 |
| 学历 | 0.951 | 3 |
| 公司所在地 | 0.959 | 1 |
| 公司性质 | 0.953 | 2 |

从上表可以看出：针对本次 4 个评价项，公司所在地的综合评价最高（关联度为：0.959），其次是公司性质（关联度为：0.953）。

## 三、描述性统计分析

本文研究数据总数为 40000 条，基于上述指标选取，本次描述性统计分析选择三个行业进行分析，即金融行业、互联网行业和生产制造行业。

### （一）金融行业

通过 Pearson 相关系数检验和灰色关联度分析，得出在金融行业，公司所在地、职位与薪资平均值关联度较大，因此做出以下分析：

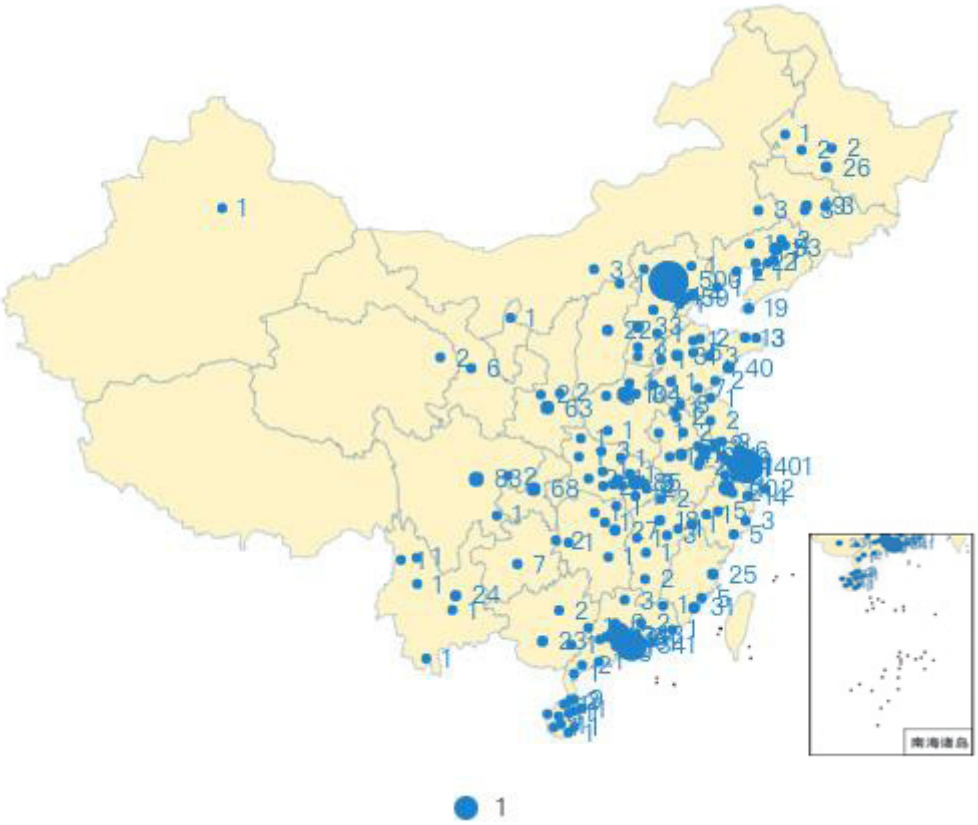



图 2　金融行业散点地图

在互联网招聘数据中，经过筛选一共筛选出 2999 条金融行业数据。由散点图呈现所示，在金融行业，招聘公司多集中于北上广地区，其次为中部地区，由此可以看出：在金融行业中，公司所在地多集中于东部发达地区。

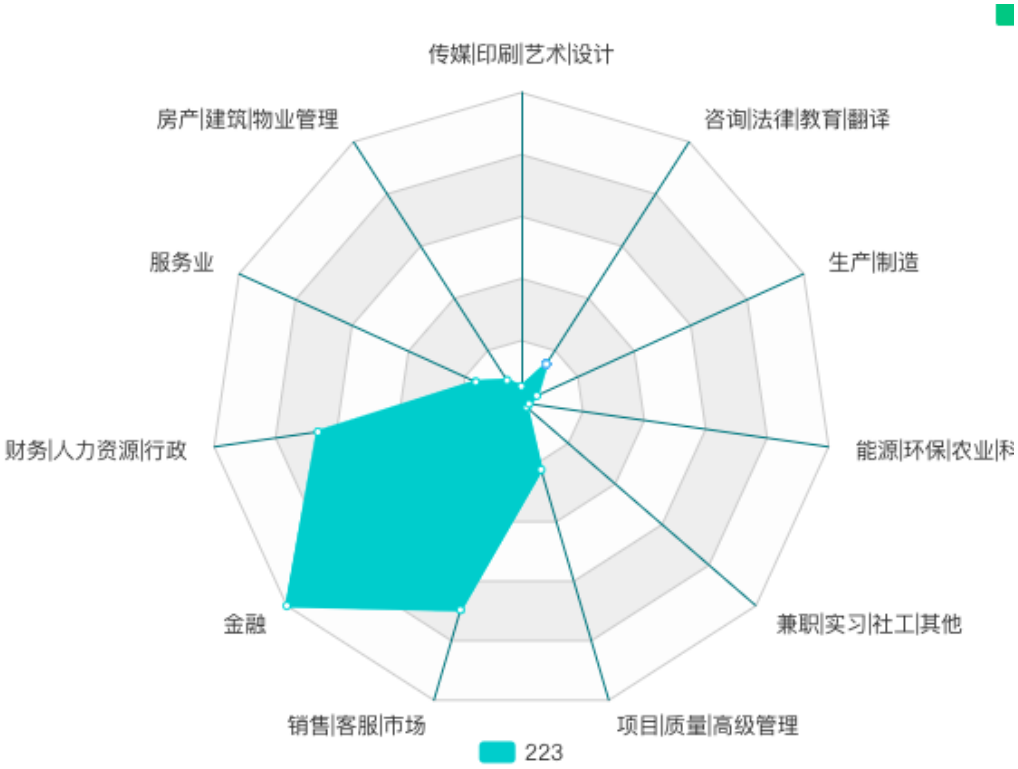



图 3　金融行业雷达图

由雷达图所示，在互联网招聘的金融行业市场中，招聘职位多为金融、销售、客服、市场职位，由此可以看出：在互联网招聘市场中，招聘者空缺的职位与其所在行业关联度较大，应聘者专业性更强。

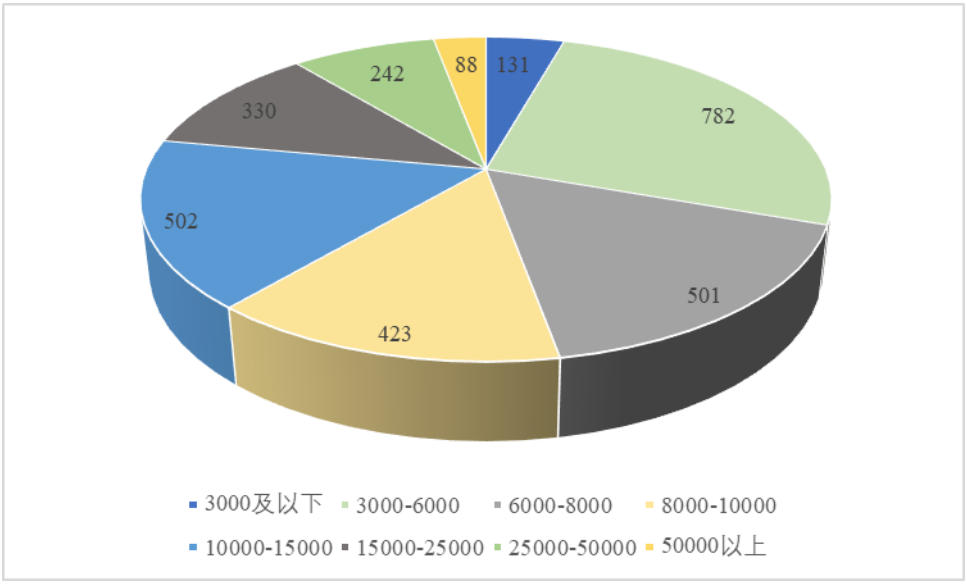



图 4　金融行业薪资 3D 饼图

由 3D 饼图显示，在金融行业中，平均薪资 3000-6000 员的岗位最多，其次是 8000-15000。由此可以看出：在金融行业互联网招聘中，作为支撑国家经济发展的重要行业之一，其薪资水平较为良好。

## （二）互联网行业

通过 Pearson 相关系数检验和灰色关联度分析，得出在互联网行业，学历、公司性质与薪资平均值关联度较大，因此做出以下分析：

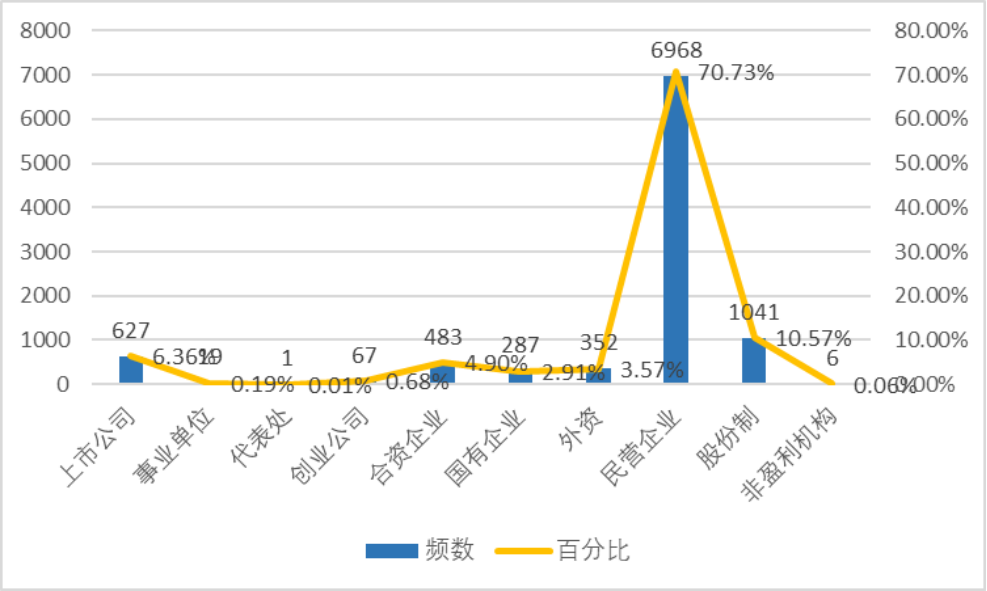



图 5　互联网行业公司性质折线图

在互联网招聘数据中，经过筛选一共筛选出 9851 条互联网行业的数据，由上图所示，在 9851 条数据中，公司性质为民营企业所招聘的人才最多，其次是上市公司、股份制公司、合资企业，由此可以看出，在互联网行业中应聘人才大多数倾向民营企业。

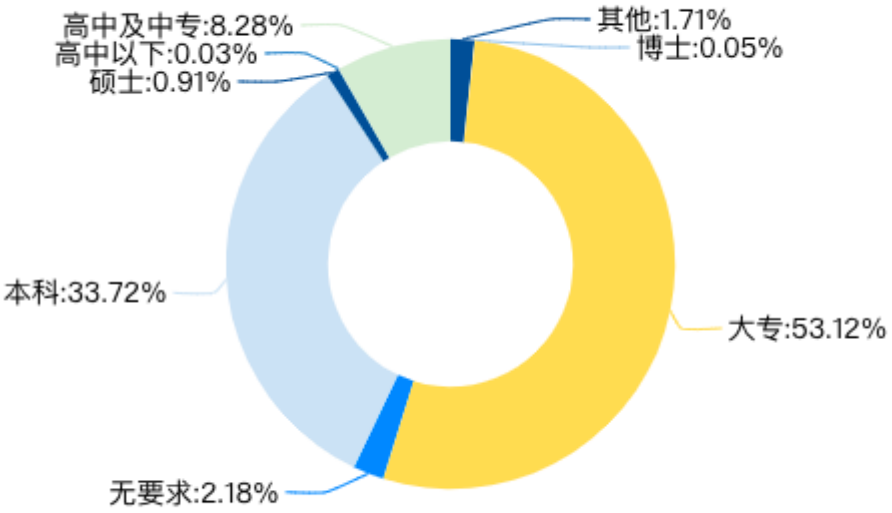



图 6　互联网行业学历环形图

由上图所示，在互联网行业中，本科及大专学历人才占比较大，无学历要求的招聘岗位较少，由此可以看出在互联网行业中招聘者对学历要求较高，应聘者中学历较高者也占绝大比例。

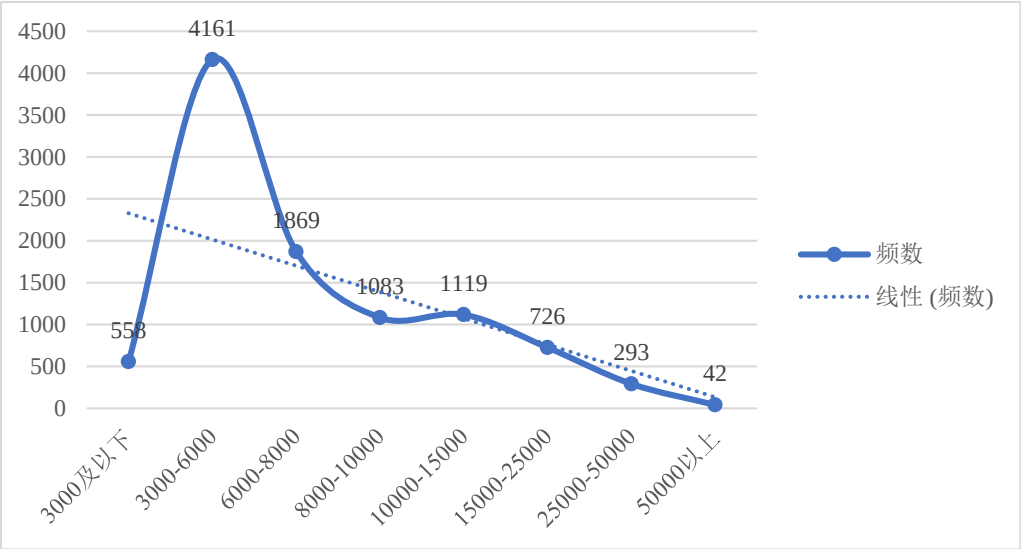



图 7　互联网行业薪资曲线图

由上述曲线图可以看出，在互联网行业中，薪资水平为 3000-6000 元的岗位占比最大，其次为 6000-8000 元薪资水平的岗位。由此呈现出：作为新兴产

业的互联网行业，其薪资水平大致与金融行业相似，区别在于其薪资水平提升空间较大。

## （三）生产制造行业

通过 Pearson 相关系数检验和灰色关联度分析，得出在生产制造行业，公司所在地、公司性质与薪资平均值关联度较大，因此做出以下分析：

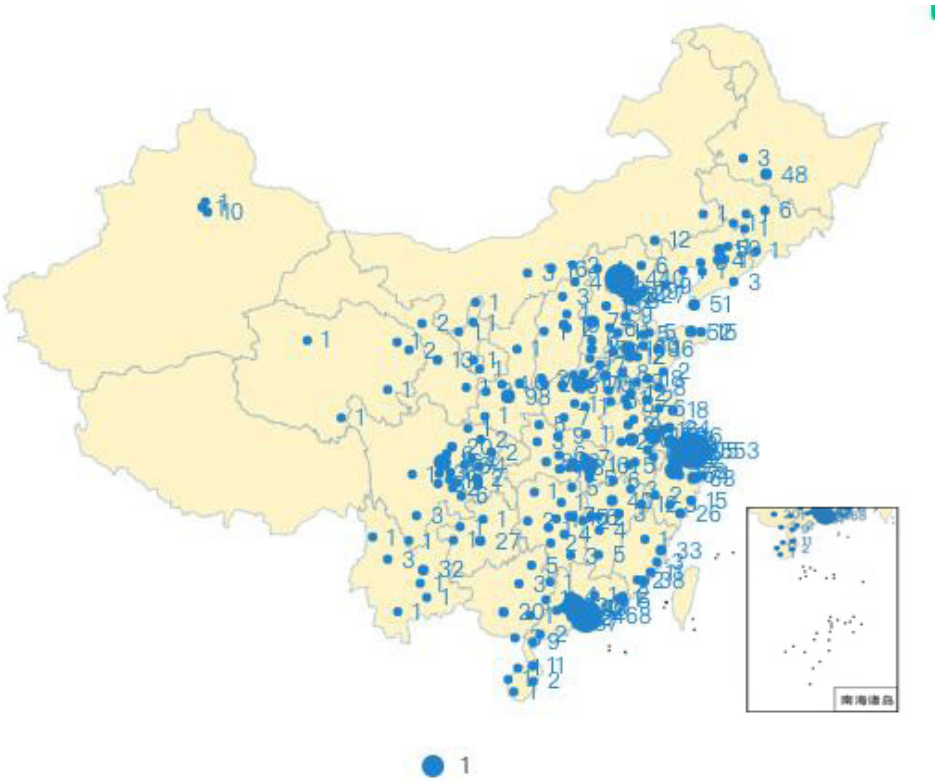



图 8　生产制造行业地域散点地图

在互联网招聘数据中，经过筛选一共筛选出 5679 条生产制造行业的数据由热力图呈现所示，在生产制造行业，招聘公司多集中于北上广地区，其次为中部地区，由此可以看出生产制造行业中，公司所在地多集中于中部、东部地区。

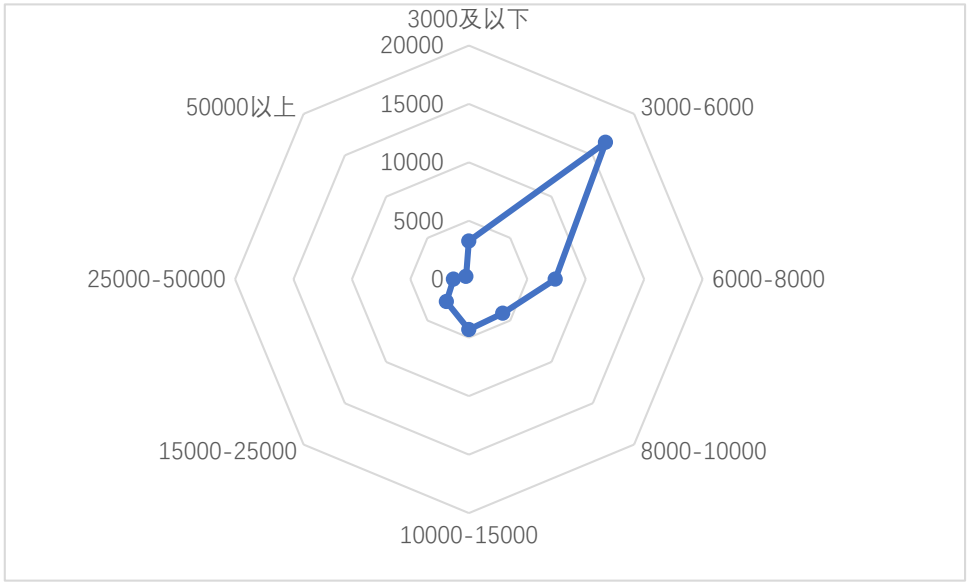



图 9　生产制造行业公司类型折线图

由上图所示，在生产制造行业中，公司性质为民营企业占比较大，超过一半，其次为股份制公司、外资和合资企业。由此可以看出在生产制造行业中招聘者所在公司性质大多为民营企业，应聘者更倾向于民营企业。

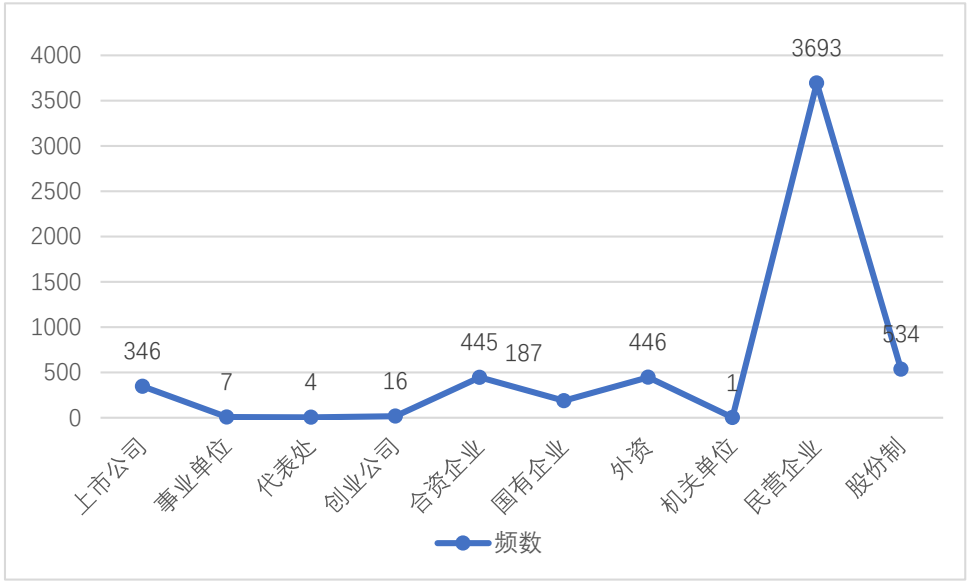



图 10　生产制造行业薪资水平雷达图

由雷达图所示，在生产制造行业中，薪资水平为 3000-6000 元的岗位占比最大，其次为 6000-8000 元薪资水平的岗位。由此呈现出，作为传统产业的生产制造行业，其薪资水平大致与新兴行业相似，区别在于其薪资水平较高的岗位数量较少。

## 四、模型建立

### （一）模型一：LightGBM 模型对浏览量特征重要值排序

LightGBM(LightGradient Boosting Machine)是一种基于梯度提升决策树（GBDT）算法的开源框架，因其高效、快速和并行的优点被广泛用于处理分类、回归等问题，能够保证较好的分类和预测结果。

1.基于 GBDT 的嵌入式特征选择

本文根据研究的互联网市场招聘对于 40000 条数据选择嵌入式特征选择方法即决策树模型和线性模型，在此基础上设计基于 GBDT 的嵌入式指标选择算法，用以计算基于 10 个指标 40000 条数据的特征重要性，在研究过程中，根据 10 个指标对与互联网招聘的浏览量的贡献，自动进行特征选择。而后进行特征进行重要性排序，丢弃不相关指标，根据上述原则，基于 GBDT 的特征选择过程设计如下：

①选择基本分类器，梯度提升决策树（GBDT）是本文的基础分类器，GBDT 的含义是一个用梯度提升策略训练的决策树模型，模型的结果是一组 CART-Tree 集合，其中，学习前几棵树预测结果的残差，$T1 \ldots\ldots Tn$模型的最终输出是每棵树中一个样本的结果之后，公式如下：

$$\bar{y} = \sum_{n=1}^{N} f_n(x), fn \subset T \tag{1}$$

fn 表示样本到树输出的映射。

②计算基于杂质的特征重要性。基于杂质的特征重要性也称为基尼重要性。通过将 GBDT 中所有树上每个特征的基尼杂质相加得到的快速浏览量可以衡量各项指标对重要性特征排序贡献。基尼重要性越高，意味着该特征越重要。

基尼重要性的计算公式如下：

$$Gini(D) = 1 - \sum_{k=1}^{|Y|} Pk^2 \quad (2)$$

其中 Pk 表示样本集中具有类的样本所占比例。

③特征重要级排序

根据浏览量各影响指标重要性的计算结果，根据诸因素对浏览量的指标按特征重要度排序。

④设置筛选阈值

设置筛选阈值。阈值是浏览量指标重要性的中位数。对重要性高于阈值的特征进行保留，对其他特征进行丢弃。

2.LightGBM 实证分析

①数据预处理

本文以三个行业中的共同影响因素：经验、职位、学历、公司所在地和薪资平均值等指标为基础，使用 leaf-wise 算法即 num_leaves 调节树的复杂程度，首先对不平衡的 40000 条数据进行数据预处理，对数据进行清洗并对指标和维度进行归一化数据处理，对哑变量进行处理后作为模型输入，以是否对互联网招聘浏览量产生影响作为模型输出，得到数据集如下所示：

表 7 LightGBM 数据集

| 特征描述 | 类型 | 个数 |
|---|---|---|
| 专业门类 | Nominal | 302 |
| 公司所在地 | Nominal | 231 |
| 薪资平均值 | Numerical | 203 |
| 一级职位 | Nominal | 195 |
| 薪资上限 | Numerical | 166 |
| 公司性质 | Nominal | 96 |
| 经验 | Numerical | 89 |
| 薪资下限 | Numerical | 88 |
| 学历 | Nominal | 70 |
| 一级行业 | Nominal | 60 |

②数据训练

利用 LightGBM 模型对输入输出的数据进行训练，建立各因素与互联网招聘浏览量的树形映射关系。输入经验、职位、学历、公司所在地等因素时，模型即对各个特征是否对互联网招聘产生影响进行重要性分析，接下来，本文运用训练数据训练 GBDT 模型 LightGBM，采用 boosting 策略来继承多个决策树，每个决策树使用损失函数的负梯度作为残差近似值来拟合新的决策树。

由于 LightGBM 中包括但不限于对模型影响较大的参数，因此本文采用网格调参，使用穷举搜索，选取每一步的迭代步长分别为 0.1、0.3、0.6。每棵树中最大叶子节点数量为 16、32、64。每棵随机采样的列数占比为为 0.5、0.8、1，控制过拟合的树的最大深度分别为-1、3、5、8，遍历循环，得出最好的参数组合是 0.8、0.1、3、16。

根据优化过后的参数得出，经验，职位，学历和公司所在地与浏览量之间的关联度较高，再根据 LightGBM 进行关于浏览量的特征重要性选择，如下图所示：

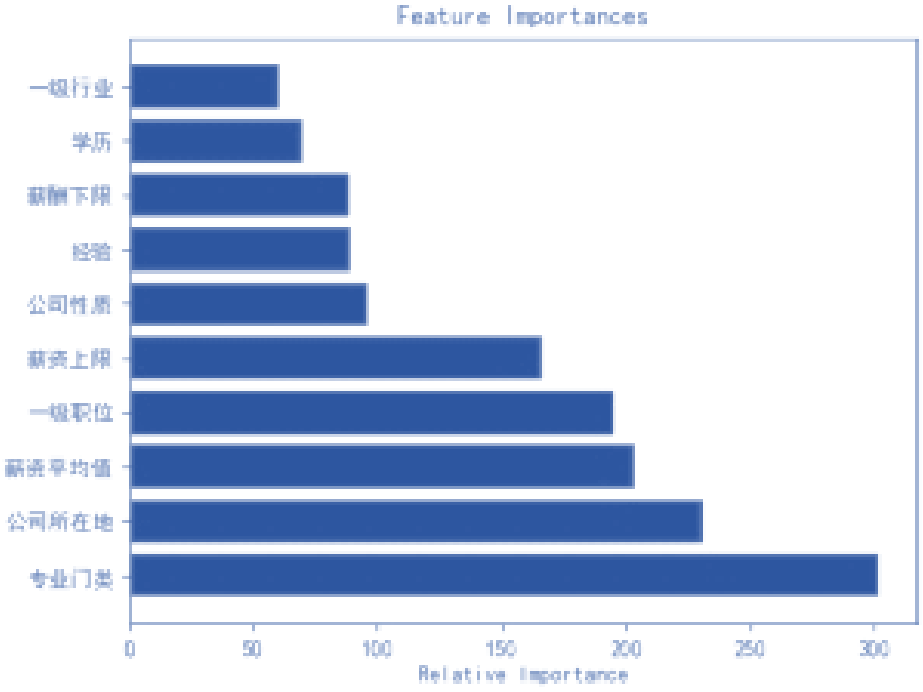



图 11LightGBM 重要性特征排序图

通过对于特征权重排序可知，专业门类、公司所在地、薪资平均值、一级职位、薪资上限维度对于互联网招聘浏览量影响较大，主要原因在于目前的就业市场企业对于专业技能的需求越来越高，而同时应聘者也希望把自己所学的知识运用于工作中，由此专业门类在互联网浏览量中所占比重较大，而公司所在地则是直接对于应聘者以后的发展起到至关重要的决定性作用，薪资平均值是应聘者考虑就业岗位的重要因素，而职位和薪资上限同时也关系着应聘者将来的发展。

3.实验结果与分析

①评价指标

针对不平衡就业数据集，本文采用查全率 Recall）、查准率（Precision）和 F-measure 作为评价方法。假设 TP 和 FP 分别表示分类的正确数和误分数，TN 和 FN 分别表示分类的正确数和误分数，F-measure 是一种衡量少数类分类性能的评价指标，只有当 Recall 和 Precision 都较高时，才能得到较好的预测结果.

②实验结果分析

本文实验环境为：python3.6、win10 系统、内存 16G、处理器 i7-7400。针对该校的就业统计数据，为了使实验结果更具客观性，采用 10 折交叉验证进行分类，与 GBDT、BalanceCascade、SMOTE-SVM、EasyEnsemble 等不平衡分类方法进行实验对比。

表 8 对比实验结果

| 算法 | Precision | Recall | F-measure |
|---|---|---|---|
| GBDT | 79.26 | 80.57 | 79.89 |
| BalanceCascade | 94.35 | 92.01 | 92.59 |
| SMOTE-SVM | 78.96 | 76.54 | 76.49 |
| EasyEnsemble | 88.89 | 89.13 | 88.34 |
| LightGBM | 91.23 | 97.56 | 93.09 |

从表 8 实验结果可知，由于数据的不平衡性和复杂性，导致上述对比算法对预测精度低于 LightGBM，虽然本文算法的整体准确率略低于 BalanceCascad

算法，但对不平衡的少数类的预测准确率较高，所以 Recall 和 F-measure 高于其他对比算法，说明了本文方法能有效识预测互联网招聘求职者的关注点，且对少数选择的互联网招聘岗位的预测精度显著提高。

此外，通过图 11 特征值权重排序，可知专业门类、公司所在地、薪资平均值、一级职位、薪资上限维度对于互联网招聘浏览量影响较大，将特征值权重高的特征作为就业因子，以此加大对互联网招聘招聘者的相关指导和帮助，进一步提高招聘满意度。

## （二）模型二：BP 神经网络模型对薪资水平的预测

1.BP 神经网络模型理论介绍

BP 神经网络，即反向传播神经网络（Back-ProPagation），是一种适应于非线性模式识别和分类预问题的人工神经网络。BP 网络有自学习、自适应特点，具有高度非线性和较强的泛化能力，具能以任意精度逼近非线性关系，在神经网络中应用广泛。它是通过样对本数据的训练，不断修正网络权值和阈值，使误差函数沿负梯度方向下降，逼近期望输出。

BP 神经网络由输入层、隐含层和输出层组成，网络选用 S 型传递函数，

$$f(x)=\frac{1}{1+e^{-x}} \tag{3}$$

通过反传误差函数

$$E=\frac{\sum_i(t_i+o_i)^2}{2} \tag{4}$$

（$t_i$为期望输出、$o_i$为网络的计算输出），不断调节网络权值和阈值使误差函数 E 达到极小。

BP 神经网络的过程分为前向传播和后向传播两个阶段：第一阶段，正向传播，即数据从输入端输入之后，沿着神经网络的指向，乘以对应的权重之后再加和，再将结果作为输入在激活函数中进行计算，将计算的结果作为输入传递给下一个节点。依次计算从输入层经过输出层到达输出层，直到得到最终的输出，完成前向传播；第二阶段，反向传播，将输出的结果与理想的输出结果进行比较，将输出结果与理想输出结果之间的误差利用网络进行反向传播。具体的过程是通过多次迭代的过程，不断地对网络上各个节点间的所有的权重进行调整，权重调整的方法采用梯度下降法。

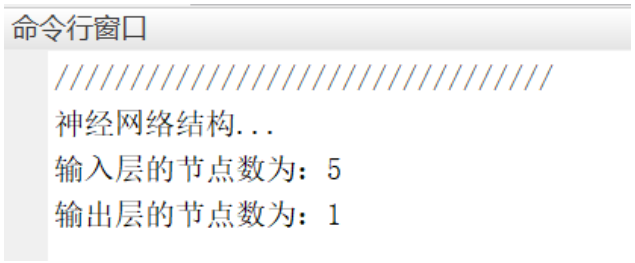
2.BP 神经网络模型实证分析

①网络结构设计

1)输入输出层的设计

通过 Pearson 相关系数检验和灰度分析得知，在金融行业、互联网行业、生产制造行业，求职者经验、求职者学历，公司所在地和职位四个指标都与对薪资平均值有显著影响。为探究上述变量与平均薪资之间的关系，我们选取行业、求职者经验、求职者学历、公司所在地、职位作为输入，以薪资平均值作为输出，所以输入层的节点数为 5，输出层的节点数为 1。

图 13 输入层与输出层节点数




代码如下：

```
%% 获取输入层节点、输出层节点个数
inputnum=size(input,2);
outputnum=size(output,2);
disp('//////////////////////////////////')
disp('神经网络结构...')
disp(['输入层的节点数为：',num2str(inputnum)])
disp(['输出层的节点数为：',num2str(outputnum)])
disp(' ')
disp('隐含层节点的确定过程...')
```

2)隐含层设计

本文采用一个隐含层的三层多输入单输出的 BP 神经网络模型。对于隐含层神经元节点数的确定，节点个数过多，会加大网络计算量并容易产生过度拟合问题；节点数个数过少，则会影响网络精度，达不到预期效果。网络中隐含层神经元的数目与实际问题的复杂程度、输入和输出层的神经元数以及对期望误差的设

定有着直接的联系。本文在确定隐含层神经元个数的问题上参照了以下的经验公式:

$$l = \sqrt{m+n} + a \tag{5}$$

其中, n 为输入层神经元个数, m 为输出层神经元个数, a 为[ 1, 10]之间的常数。

根据上式可以计算出神经元个数为 3-12 个之间，通过比较各节点数训练集的均方误差，当隐含层节点数为 9 时，均方误差最小，为 0.0014126，故在本次实验中最佳隐含层节点数为 9。

代码如下：

```
%确定隐含层节点个数
%采用经验公式 hiddennum=sqrt(m+n)+a，m 为输入层节点个数，n 为输出层节点个数，a 一般取为 1-10 之间的整数
MSE=1e+5; %初始化最小误差
transform_func={'tansig','purelin'}; %激活函数
train_func='trainlm';     %训练算法
for hiddennum=fix(sqrt(inputnum+outputnum))+1:fix(sqrt(inputnum+outputnum))
```

表 9 隐含层节点数的确定过程

| 隐含节点数 | 训练集均方误差 |
|---|---|
| 3 | 0.0014902 |
| 4 | 0.0014323 |
| 5 | 0.001458 |
| 6 | 0.0014813 |
| 7 | 0.0014742 |
| 8 | 0.0014316 |
| 9 | 0.0014126 |
| 10 | 0.0014513 |
| 11 | 0.001465 |
| 12 | 0.0014675 |

3）激活函数的选取

BP 神经网络的激活函数通常采用 Sigmoid 函数、Softmax 和 ReLU(Rectified Linar Unit)修正线性单元函数等。本文选择 S 型正切函数 tansig 作为隐含层神经元的激活函数。而由于网络的输出归一到[ -1, 1]范围内，因此预测模型选取 S 型对数函数 tansig 作为输出层神经元的激活函数。

本文构建的 BP 神经网络模型如图所示：

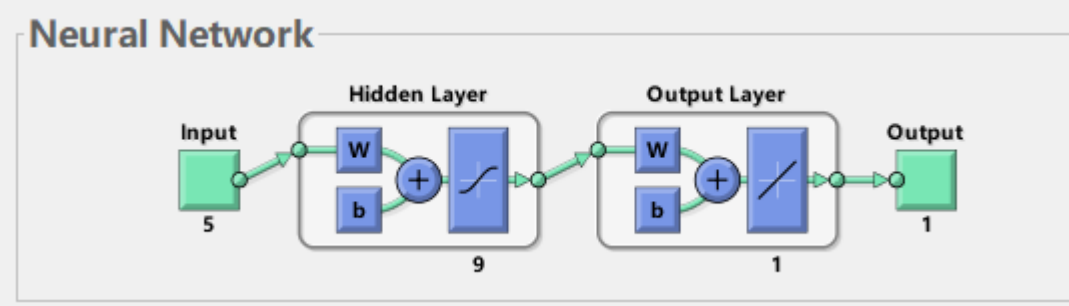



图 14BP 神经网络模型

②实验结果分析

本文选取有关金融、互联网、生产制造行业，求职者经验，求职者学历，公司所在地，职位和平均薪资水平六个指标 18528 条数据进行实验分析，其中测试样本数目为 50 条。将训练样本数据归一后输入网络，确定网络参数配置，其中网络迭代次数 epochs 为 1000 次，学习速率 lr 为 0.01，训练目标最小误差 goal 为 0.000001。通过训练样本数据对 BP 神经网络进行训练，确定输入节点数为 5，输出节点数为 1，隐含层节点为 9。随后用训练好的模型进行仿真，进行预测结果反归一化与误差计算，结果如图所示：

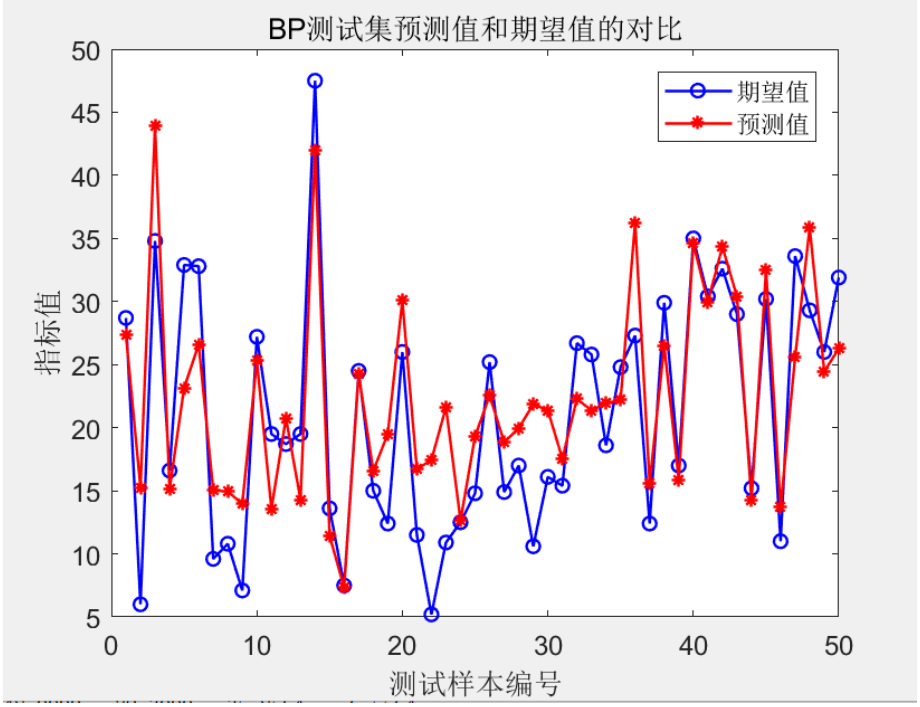



图 15 预测值和真实值的分析图像

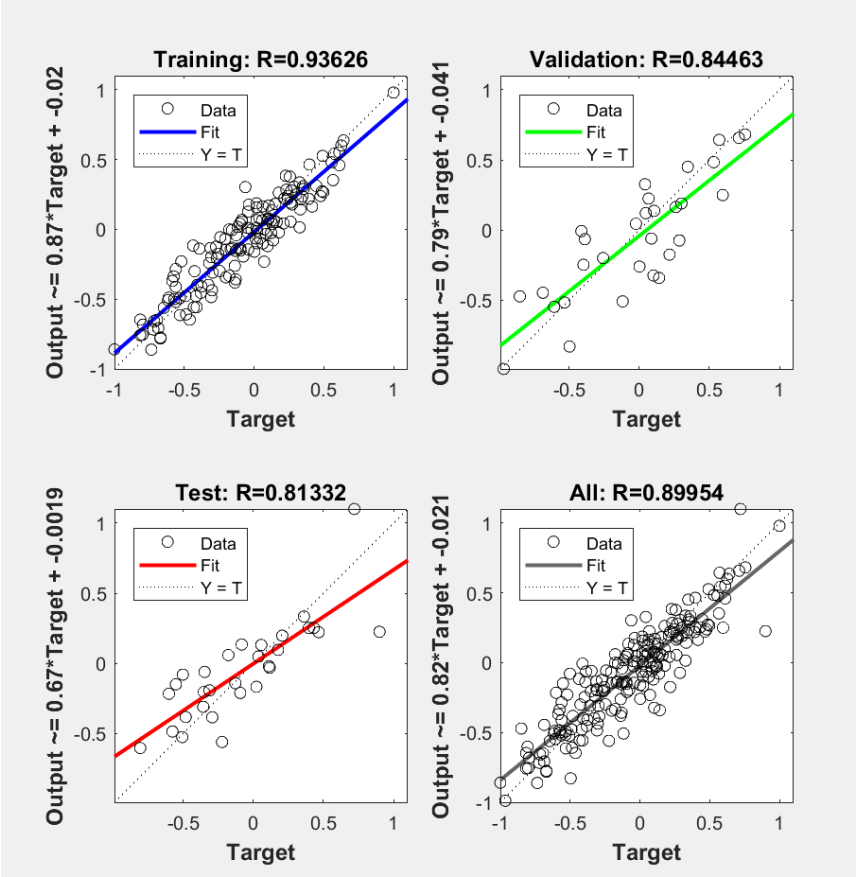



图 16 各个样本集和总体的相关性分析

如图所示，样本真实值和预测值的误差较小，且 R 值较大，说明说明模型较好，预测精度高。通过此模型，在给出行业。求职者经验、求职者学历、 公司所在地、职位时，可对薪资水平进行预测，具有现实参考意义。

代码如下：

```
%% 网络测试
an=sim(net,inputn_test); %用训练好的模型进行仿真
test_simu=mapminmax('reverse',an,outputps); % 预测结果反归一化
error=test_simu-output_test;          %预测值和真实值的误差
%%真实值与预测值误差比较
figure
plot(output_test,'bo-','linewidth',1.2)
hold on
plot(test_simu,'r*-','linewidth',1.2)
legend('期望值','预测值')
xlabel('测试样本编号'),ylabel('指标值')
title('BP 测试集预测值和期望值的对比')
set(gca,'fontsize',12)
igure
plot(error,'ro-','linewidth',1.2)
xlabel('测试样本编号'),ylabel('预测偏差')
title('BP 神经网络测试集的预测误差')
set(gca,'fontsize',12)
%计算误差
[~,len]=size(output_test);
```

```
SSE1=sum(error.^2);
MAE1=sum(abs(error))/len;
MSE1=error*error'/len;
RMSE1=MSE1^(1/2);
MAPE1=mean(abs(error./output_test));
r=corrcoef(output_test,test_simu);      %corrcoef 计算相关系数矩阵,包括自相关和互相关系数
R1=r(1,2);
```

3.BP 神经网络模型与多元线性回归模型对比分析

①多元线性回归模型

多元线性回归模型是描述一个因变量 Y 与一个或多个自变量 X 之间的线性依存关系，用一定的线性拟合因变量和自变量的关系,确定模型参数来得到回归方程,并用回归方程预测因变量的变化趋势,运用回归分析方法能够建立反映具体数量关系的数学模型,即回归模型。方程如下所示：

$$Y_{hat} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_n X_n + e \quad (6)$$

其中，n 是解释变量的个数，$\beta_i$是常数项，$\beta_1$～$\beta_n$为回归系数，e 是误差项的随机变量值，是去除 n 个自变量对$Y_{hat}$影响后的随机误差。

实验以求职者学历，求职者经验，职位和公司所在地作为输入参数，平均薪资水平为输出参数，采用多元线性回归模型对分别金融行业、互联网行业、生产制造行业平均薪资水平进行预测。

表 10 独立性检验

| | R | 德宾-沃森 |
|---|---|---|
| 金融行业 | .652a | 1.992 |
| 加工制造行业 | .624a | 1.969 |
| 互联网行业 | .695a | 1.982 |

由表可知，三行业分析中 DW 值接近 2，样本独立，R 值大于 0.5，可用于多元线性回归模型分析。具体模型参数如下：

表 11 金融行业标准化系数表

| | | | 标准化系数 Beta | t | 显著性 |
|---|---|---|---|---|---|
| (常量) | 1917.125 | 1508.942 | | 1.271 | 0.204 |
| 公司所在地 | 2300.570 | 397.745 | -0.100 | -5.784 | 0.000 |
| 经验 | 4771.943 | 306.000 | 0.288 | 15.595 | 0.000 |
| 学历 | 1511.612 | 484.984 | 0.058 | 3.117 | 0.002 |
| 职位 | 483.303 | 122.503 | 0.068 | 3.945 | 0.000 |

由系数表可知，在金融行业中，公司所在地、经验、学历、职位对薪资有显著性影响。

表 12 加工制造业标准化系数表

| | | | 标准化系数 Beta | t | 显著性 |
|---|---|---|---|---|---|
| (常量) | -392.157 | 461.900 | | -0.849 | 0.396 |
| 经验 | 2447.316 | 91.693 | 0.323 | 26.690 | 0.000 |
| 职位 | 22.585 | 36.773 | 0.007 | 0.614 | 0.539 |
| 学历 | 2247.663 | 133.940 | 0.205 | 16.781 | 0.000 |
| 公司所在地 | -830.168 | 103.848 | -0.097 | -7.994 | 0.000 |

由系数表可知，在生产制造行业中，公司所在地、经验、学历对薪资有显著性影响。

表 13 互联网行业标准化系数表

| | | | 标准化系数 Beta | t | 显著性 |
|---|---|---|---|---|---|
| (常量) | 998.497 | 283.052 | | 3.528 | 0.000 |
| 学历 | 2414.254 | 94.885 | 0.234 | 25.444 | 0.000 |
| 经验 | 2142.949 | 58.861 | 0.335 | 36.407 | 0.000 |
| 职位 | 39.410 | 23.362 | 0.015 | 1.687 | 0.092 |
| 公司所在地 | 993.142 | 72.577 | -0.121 | -13.684 | 0.000 |

由系数表可知，在互联网行业中，公司所在地、经验、学历对薪资有显著性影响。

②模型对比分析

通过两种方法对平均薪资水平进行预测后，对于 R 值的对比，可以看出 BP 神经网络在进行预测时拟合度更高。由于因变量和自变量间呈现的是非典型的线性相关关系，多元线性回归无法准确反映他们之间的关系，而 BP 神经网络可以处理复杂空间的非线性系统，通过调整内部的权重使网络优化效果提高，通过数据训练达到一个稳定输出，使得其比多元线性回归模型在平均薪酬水平预测时具有更高的精度和应用范围。

## 五、结论与建议

### （一）结论

本文通过回顾我国就业政策以及目前我国的就业背景，深度剖析了互联网招聘市场招聘者和应聘者需求不对称之间的难题。在此基础上，选取新兴行业——互联网行业、国家经济发展重要支撑——金融行业、待转型升级行业——生产制造行业等三个行业作数据分析，运用 Pearson 相关系数检验分析三个行业中影响薪资的诸因素，并运用灰色关联度分析法进一步分析诸因素的影响大小，其次运用 LightGBM 模型和 BP 神经网络分别研究薪资预测与浏览量，所得研究结果如下：

1.招聘者：LightGBM 模型

根据 LightGBM 模型特征重要性分析可知，专业门类、公司所在地、薪资平均值、职位、薪资上限对于互联网招聘浏览量影响较大。因此得出位于北上广等城市、专业性强的岗位、薪资高的职业在招聘市场中更具优势。

2.求职者:薪资

2.1 金融行业薪资水平

在金融行业中，影响薪资因素为公司所在地和职位，因此得出支撑国家经济发展的金融行业越集中于经济发达地区，其薪资水平越高；

2.2.互联网行业薪资水平

在互联网行业中，影响薪资水平的因素为学历和公司性质，因此得出学历越高的应聘者所得薪资越高，民营企业薪资水平高于其他企业；

2.3.生产制造行业薪资水平

在生产制造行业中，影响薪资水平的因素为公司所在地和公司性质，因此得出在加工制造行业中公司所在地的工业基础条件越好，其薪资水平越高，民营企业薪资水平高于其他企业。

## （二）建议

1.企业招聘：根据岗位浏览量合理设置招聘要求

根据 LightGBM 模型的特征重要性分析，将特征值权重高的特征作为就业因子，以此加大对互联网招聘招聘者的相关指导和帮助，进一步提高招聘满意度。

招聘者应把相对重要性不高的指标进行合理设置。把招聘专业卡的过于严格将会带来很大程度浏览量的下降，但如果将应聘者学历和工作经验设置的要求过高会使浏览量有一定程度的下降，招聘者应根据自身需求合理设置人才招聘条件。

2.求职应聘：根据显著因子合理考虑就业岗位

2.1 金融行业：根据学历因素合理考虑就业地域

应聘者应根据金融行业对于学历的重视程度进行自身素质的评估，并综合考虑经济发达地区的金融就业岗位，若自身条件尚不达标，可考虑其他薪资相对不高但前景好的二三线城市。

2.2 互联网行业：根据学历因素合理考虑公司性质

应聘者应根据互联网行业对于学历的要求进行自身技术的评估并综合考虑不同性质的公司的互联网行业招聘岗位。若自身条件尚不达标，可考虑竞争压力相对不大但薪资较高的上市公司。

2.3 生产制造行业：根据公司所在地合理考虑公司性质

应聘者应根据生产制造行业技术水平要求不高所以对于地域较为追求的特点，综合考虑公司性质，对于数据新动能转型升级的行业发展前景进行评估，选择自身最合适的岗位。

# 参考文献

[1]白尧•基于“互联网+”时代的员工招聘管理路径研究[J]•中国管理信息化,2019,22(22):67-68.

[2]李燕萍，齐伶圆• “互联网+”时代的员工招聘管理:途径、影响和趋势[J]•中国人力资源开发,2016(18):6-13+19.

[3]张博，杨婷婷，韩飞•互联网时代下多重互动式社会化网络招聘模式研究——以猎聘网为案例[J]•中国人力资源开发,2016(18):20-25.

[4]谷彬•互联网大数据与人才精细化管理[J]•调研世界,2016(09):50-53.

[5]徐汝婷，蔡晓晶•“互联网+”时代的员工招聘[J]•商,2015(42):37+32.

[6]黄敬宝•“互联网+”时代的青年就业与新思维[J]•中国青年社会科学,2015,34(05):43-49.

[7]侯艺•保就业背景下青年就业现状研究[J]•中国青年研究,2020(09):107-112.

[8]陈有华，张壮•新冠肺炎疫情认知对就业预期的影响[J]•华南农业大学学报(社会科学版),2020,19(04):105-119.

[9]王祎•互联网背景下腾讯公司的招聘渠道研究[J]•商讯,2019(24):13-14.

[10]郭睿•学历、专业错配与高校毕业生就业质量[D]•湖南大学,2019.

[11]Celia P, Kaplan Adam, Siegel Yan, Leykin, Nynikka R, Palmer Hala，Borno Jessica Bielenberg, Jennifer Livaudais-Toman, Charles Ryan, Eric J, Small. A bilingual, Internet-based, targeted advertising campaign for prostate cancer clinical trials: Assessing the feasibility, acceptability, and efficacy of a novel recruitment strategy[J]• Contemporary Clinical Trials Communications,2018,12.

[12]Kristin L, Corey. Mary K, McCurry, Kristen A, Sethares, Meg Bourbonniere, Karen B, Hirschman, Salimah H • MeghaniUtilizing Internet-based recruitment and data collection to access different age groups of former family caregivers[J] • Applied Nursing Research,2018,44.

# 附录

1.BP 神经网络代码

```
%%  初始化
clear
close all
clc
format short
%%  读取读取
data=xlsread('数据总.xlsx','Sheet1','A1:F18528'); %%使用 xlsread 函数读取 EXCEL
中对应范围的数据即可
%输入输出数据
input=data(:,1:end-1);       %data 的第一列-倒数第二列为特征指标
output=data(:,end);   %data 的最后面一列为输出的指标值
N=length(output);     %全部样本数目
testNum=50;     %设定测试样本数目
trainNum=N-testNum;       %计算训练样本数目
%%  划分训练集、测试集
input_train = input(1:trainNum,:)';
output_train =output(1:trainNum)';
input_test =input(trainNum+1:trainNum+testNum,:)';
output_test =output(trainNum+1:trainNum+testNum)';
%%  数据归一化
[inputn,inputps]=mapminmax(input_train,0,1);
[outputn,outputps]=mapminmax(output_train);
```

```
inputn_test=mapminmax('apply',input_test,inputps);
%% 获取输入层节点、输出层节点个数
inputnum=size(input,2);
outputnum=size(output,2);
disp('/////////////////////////////////')
disp('神经网络结构...')
disp(['输入层的节点数为：',num2str(inputnum)])
disp(['输出层的节点数为：',num2str(outputnum)])
disp(' ')
disp('隐含层节点的确定过程...')
%确定隐含层节点个数
%采用经验公式 hiddennum=sqrt(m+n)+a，m 为输入层节点个数，n 为输出层节点
个数，a 一般取为 1-10 之间的整数
MSE=1e+5; %初始化最小误差
transform_func={'tansig','purelin'}; %激活函数
train_func='trainlm';     %训练算法
for
hiddennum=fix(sqrt(inputnum+outputnum))+1:fix(sqrt(inputnum+outputnum))+10
%构建网络
net=newff(inputn,outputn,hiddennum,transform_func,train_func);
% 网络参数
net.trainParam.epochs=1000;              % 训练次数
net.trainParam.lr=0.01;                          % 学习速率
net.trainParam.goal=0.000001;            % 训练目标最小误差
```

```
% 网络训练
net=train(net,inputn,outputn);
an0=sim(net,inputn);  %仿真结果
mse0=mse(outputn,an0);  %仿真的均方误差
disp(['隐含层节点数为',num2str(hiddennum),'时，训练集的均方误差为：
',num2str(mse0)])
%更新最佳的隐含层节点
if mse0<MSE
MSE=mse0;
hiddennum_best=hiddennum;
end
end
disp(['最佳的隐含层节点数为：',num2str(hiddennum_best),'，相应的均方误差为：
',num2str(MSE)])
%% 构建最佳隐含层节点的 BP 神经网络
net=newff(inputn,outputn,hiddennum_best,transform_func,train_func);
% 网络参数
net.trainParam.epochs=1000;          % 训练次数
net.trainParam.lr=0.01;              % 学习速率
net.trainParam.goal=0.000001;        % 训练目标最小误差
%% 网络训练
net=train(net,inputn,outputn);
%% 网络测试
an=sim(net,inputn_test); %用训练好的模型进行仿真
```

```
test_simu=mapminmax('reverse',an,outputps); %  预测结果反归一化
error=test_simu-output_test;          %预测值和真实值的误差
%%真实值与预测值误差比较
figure
plot(output_test,'bo-','linewidth',1.2)
hold on
plot(test_simu,'r*-','linewidth',1.2)
legend('期望值','预测值')
xlabel('测试样本编号'),ylabel('指标值')
title('BP 测试集预测值和期望值的对比')
set(gca,'fontsize',12)
figure
plot(error,'ro-','linewidth',1.2)
xlabel('测试样本编号'),ylabel('预测偏差')
title('BP 神经网络测试集的预测误差')
set(gca,'fontsize',12)
%计算误差
[~,len]=size(output_test);
SSE1=sum(error.^2);
MAE1=sum(abs(error))/len;
MSE1=error*error'/len;
RMSE1=MSE1^(1/2);
MAPE1=mean(abs(error./output_test));
r=corrcoef(output_test,test_simu);      %corrcoef 计算相关系数矩阵，包括自相关和
```

```
互相关系数
R1=r(1,2);
disp(' ')
disp('//////////////////////////////////')
disp('预测误差分析...')
disp(['误差平方和 SSE 为：                    ',num2str(SSE1)])
disp(['平均绝对误差 MAE 为：            ',num2str(MAE1)])
disp(['均方误差 MSE 为：                    ',num2str(MSE1)])
disp(['均方根误差 RMSE 为：              ',num2str(RMSE1)])
disp(['平均百分比误差 MAPE 为：  ',num2str(MAPE1*100),'%'])
disp(['相关系数 R 为：                              ',num2str(R1)])
%打印结果
disp(' ')
disp('//////////////////////////////////')
disp('打印测试集预测结果...')
disp(['      编号              实际值              预测值              误差'])
for i=1:len
disp([i,output_test(i),test_simu(i),error(i)])
end
```

# 致谢

时光荏苒，回首收到官方发布比赛文件已经历时两个半多月，到了提交论文的最后阶段。在这段充满奋斗的过程中，带给我们的学生生涯中无限的激情和收获。强烈感谢指导老师在此过程中多方面提供的支持和帮助，如果没有他们无私的帮助对我们的论文进行多次的指导和修改，我们将难以如此圆满的完成此论文。感谢本论文引用的参考资料的作者，是他们先前的经验和成果，为我们的论文提供多样的思路启发。在此再次对为我们提供帮助和支持的老师、作者和同学表示由衷的感谢！金无足赤，人无完人，由于我们学术能力有待提高，所写论文难免存在不足之处，恳请各位专家老师和同学提供批评指正！