****

**2020广州市英才计划研究报告**

**Report on the Talent Plan of Guangzhou in 2020**

|  |  |
| --- | --- |
| 题目  Title | 基于网络爬虫的  数据分析行业薪资预测 |
| 实验室  Laboratory | 第九实验室 |
| 指导老师  Major | 黎培兴 副教授 |
| 学员姓名  Student Name | 广州中学 杨锦航  广州市天河外国语学校 刘骐铭  广州市第一一三中学 张启喆  广州市第六中学 陈光裕  中山大学附属中学 张曦元  中山大学数学学院 刘广杰 |

时间： 2021 年 7 月 26 日

摘 要

人工智能的时代是用数据说话的时代，也是一个依靠数据竞争的时代。新时代，也是大数据的时代，数据分析师的意义愈发重要，各类相关职位的数量不断增加，因此，本文采用神经网络算法从薪酬职位数据中提取特征，完成数据分析行业薪酬分析与预测，具有很高的应用价值，能够带来可观的经济效益，并将进一步提高薪资市场的透明度。

本文的主要工作是研究机器学习技术在薪资预测方面的应用，通过机器学习技术用不同综合指标及各因素的分析对数据分析岗位进行薪资预测，为用户提供更准确的建议和信息，使求职者能够从海量的职位信息中选择更合适的职位，提升用户体验。

为了优化薪资预测工作，提高工作效率与预测精准度，通过引入机器学习及神经网络来实现薪资预测系统的功能，可以减少业务工作量与人工计算量，从而从较大程度上避免人工误差与用人成本高昂效率低下问题。我们通过将机器学习技术应用到薪资预测系统上，设计并训练构建实现了薪资预测神经网络模型。该模型使用了 BP 神经网络实现，基于 ReLU 激活函数，在隐藏层数量测试与避免过拟合优化中，通过设计了一系列实验，将实现的模型与同一过程不同隐藏层数量的模型相比较，我们最终选取使用 13 个隐藏层作为模型隐藏层数量。该模型在测试实验中表现较为优秀，说明了它们具有一定的实际应用参考价值，也进一步说明了机器学习在薪资预测中的应用是具有相当大的发展潜力的。

**关键词：** 薪资预测 数据分析 机器学习 网络爬虫 神经网络

ABSTRACT

The era of artificial intelligence is not only an era of speaking with data, but also an era of relying on data competition. The new era is also the era of big data. The significance of data analysts is becoming more and more important, and the number of various related positions is increasing. Therefore, this paper uses neural network algorithm to extract features from salary position data and complete data analysis, industry salary analysis and prediction, which has high application value and can bring considerable economic benefits, And will further enhance the transparency of the salary market.

The main work of this paper is to study the application of machine learning technology in salary prediction. Through the analysis of different comprehensive indicators and factors, this paper forecasts the salary of data analysis posts through machine learning technology, so as to provide users with more accurate suggestions and information, so that job seekers can choose more appropriate positions from a large amount of position information and improve the user experience.

In order to optimize the salary forecasting work and improve the work efficiency and forecasting accuracy, the function of the salary forecasting system is realized by introducing machine learning and neural network, which can reduce the business workload and manual calculation, so as to avoid the problems of manual error, high employment cost and low efficiency to a great extent. By applying machine learning technology to salary prediction system, we design and train the salary prediction neural network model. The model is implemented by BP neural network. Based on ReLU activation function, in the number test of hidden layers and avoiding over fitting optimization, a series of experiments are designed to compare the implemented model with the model with different number of hidden layers in the same process. Finally, we select 13 hidden layers as the number of hidden layers in the model. The model performs well in the test experiment, which shows that they have certain practical application reference value, and further shows that the application of machine learning in salary prediction has considerable development potential.

**KEY WORDS:** salary prediction data analysis machine learning web crawler neural network

目 录

[摘 要 i](#_Toc78188478)

[ABSTRACT ii](#_Toc78188479)

[目 录 iii](#_Toc78188480)

[第一章、绪论 1](#_Toc78188481)

[1.1 研究背景和意义 1](#_Toc78188482)

[1.1.1 研究背景 1](#_Toc78188483)

[1.1.2 意义 2](#_Toc78188484)

[1.2 国内外研究现状 2](#_Toc78188485)

[1.3 研究内容与方法 3](#_Toc78188486)

[第二章、相关理论概述 5](#_Toc78188487)

[2.1 网络爬虫介绍 5](#_Toc78188488)

[2.2 随机森林模型介绍 5](#_Toc78188489)

[2.3 神经网络模型介绍 8](#_Toc78188490)

[第三章、数据爬取与预处理 10](#_Toc78188491)

[3.1 数据分析岗定义与薪资数据的选取和获取 10](#_Toc78188492)

[3.1.1 数据分析岗的定义 10](#_Toc78188493)

[3.1.2 数据分析岗数据的选取 10](#_Toc78188494)

[3.1.3 数据分析岗薪资数据的获取 11](#_Toc78188495)

[3.2 热门城市基本指数的获取与挑选 12](#_Toc78188496)

[3.2.1 城市基本指数 12](#_Toc78188497)

[3.2.2 热门城市基本指数的挑选与获取 13](#_Toc78188498)

[3.3 数据分析岗薪资数据预处理 13](#_Toc78188499)

[3.3.1 岗位数据预处理 13](#_Toc78188500)

[3.3.2 岗位数据清洗 14](#_Toc78188501)

[3.4 热门城市综合指标的预处理 16](#_Toc78188502)

[3.4.1 缺失值处理 16](#_Toc78188503)

[3.4.2 异常值处理 16](#_Toc78188504)

[3.4.3 特征筛选 17](#_Toc78188505)

[3.4.4 小结 18](#_Toc78188506)

[3.5 数据分析岗基本特征分析 19](#_Toc78188507)

[第四章、基于神经网络的数据分析岗薪资预测 22](#_Toc78188508)

[4.1 基于随机森林算法的特征选择 22](#_Toc78188509)

[4.1.1 基于随机缺失森林的热门城市综合指标填充 22](#_Toc78188510)

[4.1.2 招聘数据集的规范化 23](#_Toc78188511)

[4.1.3 基于随机森林的训练数据特征选择 25](#_Toc78188512)

[4.2 薪资预测模型的选择 28](#_Toc78188513)

[4.2.1 回归模型 28](#_Toc78188514)

[4.2.2 模型比较 29](#_Toc78188515)

[4.3 神经网络模型的预训练 29](#_Toc78188516)

[4.3.1 训练集与测试集的划分 29](#_Toc78188517)

[4.3.2 数据分析岗薪资预测神经网络模型的预训练 30](#_Toc78188518)

[4.4 构建基于城市综合指标的数据分析岗薪资预测模型 32](#_Toc78188519)

[4.5 小结 32](#_Toc78188520)

[第五章、总结与展望 33](#_Toc78188521)

[5.1 主要工作总结 33](#_Toc78188522)

[5.2 未来工作展望 33](#_Toc78188523)

[参考文献 35](#_Toc78188524)

[致谢 36](#_Toc78188525)

[附录 37](#_Toc78188526)

第一章、绪论

1.1 研究背景和意义

1.1.1 研究背景

人工智能的时代是用数据说话的时代，也是一个依靠数据竞争的时代。随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生和所需的数据呈爆炸性增长。

90% 以上的世界 500 强企业都建立了数据分析部门。IBM、谷歌、微软等知名公司都在积极投资数据分析领域，建立数据分析相关部门，培养数据分析团队。大数据时代，不少领域不少行业的数据很容易达到数百 TB 甚至数十到数百 PB 的规模，已经远远超过了现有传统计算技术和信息系统的处理能力。因此，有必要寻求有效的大数据处理技术、方法和手段已成为现实世界中的迫切需求。

正是在这样的背景下，数据分析师成为了热门抢手的职业，从业人数的激增和技术的进步导致了这个行业竞争压力非常大，而从业人员也在复杂的就业选择中表现得十分茫然 —— 我该选择什么公司？这些公司有哪些有适合我的岗位？我想应聘的那些岗位，对我自身的学历和工作经验又有怎么样的要求？这个岗位有没有双休或者年假？我在这里工作能拿到多少薪水和奖金？往往一大堆的疑问总会回到岗位的薪资这个问题上来，归根结底，人们在就业时往往会对获得的薪资水平比较在意。

基于这样的现状，本文旨在通过以机器学习为主的技术构造出一个适用于数据分析岗位就业分析的人工神经网络模型，通过对中国数据分析师薪资情况的分析，对未来计划在数据分析行业或其他方面就业的人群提供分析指导，为“就业难”的数据分析行业从业人员助一臂之力。

1.1.2 意义

目前，虽然在数据分析领域已经有一些工作薪酬领域的研究，但是这方面的研究还不够深入。同时，神经网络的蓬勃发展为人工智能的发展带来了机遇，也为薪酬预测领域的研究带来了新思路。目前神经网络乃至深度学习研究领域的不断突破是通过使用不同的激活函数来消除梯度耗散的问题。近年来，神经网络在图像分类、语音识别等领域取得重大突破，获得了优于传统机器学习模型的效果。

与传统的薪资分析预测方法相比，本文将使用随机森林、BP神经网络等机器学习算法从薪资数据中提取特征，挖掘文本数据中的深层特征，并与传统机器学习模型进行比较。整合薪酬数据的特点，建立不同的模型对职位招聘信息数据进行分析，提高模型的预测准确率，完成对数据分析行业职位薪酬的预测，并取得一定的指导作用。

薪酬研究和预测的结果越准确，可以为用户提供越精准的建议和信息，使求职者能够从海量的职位信息中选择更合适的职位，提升用户体验，这在当今这个信息爆炸的时代是非常重要。因此，薪酬预测领域的研究充满前景。

综上，本文的研究内容不仅是一项具有挑战性的研究，而且具有很高的应用价值，能够带来可观的经济效益，并将进一步提高薪酬市场的透明度。

1.2 国内外研究现状

在薪水预测这个研究领域，国外研究人员的研究取得的成果较多。虽然很长一段时间比较沉寂，但是最近几年，国内外研究人员对于这方面的研究关注开始逐渐增多。早在 2012 年，Adzuna 公司就在 Kaggle 平台上举办了关于职位薪水预测研究的比赛图片，比赛的研究内容就是想要通过实现一个有效的算法模型对于英国不同的职位薪水进行分析预测，根据招聘广告的薪水建立一个预测引擎，同时可以进一步改善求职者找工作时的体验，帮助招聘人员和求职者明白不同职位的市场价值，为这个重要的市场带来更多的透明度。

国外的研究人员对于薪水预测的研究主要集中于通过不同的机器学习模型,同时利用特征工程的手段选取效果良好的特征，建立不同的薪水预测模型，完成对于垂直行业不同职位的薪水预测。日本 BizReach 人工智能技术集团研究员图片，实现了双向 GRU-CNN 模型，以预测预期收入。据观察，他们提出的模型优于除 TextCNN、RCNN、BidLSTM 和 ResNet 之外的现有其他模型，并结合深度网络完成了工资预测文本回归任务。IEEE 大会上来自泰国的研究员图片在性能方面将随机森林和梯度提升树等相关算法进行了比较，比较后将特征选择方法应用于深度学习，完成了对泰国 170 万用户的薪资预测。Shaun Jackman图片等人利用套索回归以及岭回归还有神经网络模型获取职位数据中的薪水特征，通过实验对比不同模型对于薪水预测的效果。

早期，国内对于薪资预测方面的相关研究进行的较少，大部分的数据与数据分析都是基于调查问卷与其他现场调研方式来获得求职者相关信息和求职结果，也有通过对于国外的人力资源管理中的薪酬理论的学习和借鉴，借助经济学的方式对于职位的市场价值进行预估。但是近些年，国内研究人员对于薪水预测这一领域的研究也逐渐活跃起来。潘博等人图片利用 Doc2vec 来处理可变长的非结构化薪水文本数据，提取文本的特征，更为全面的理解文本的语义特征，计算每个职位的文本信息特征向量，再结合随机森林，支持向量机等机器学习算法建立薪水预测模型。谷承维图片基于深度学习，对传统机器学习模型进行研究，提出两种全新模型与传统特征融合的薪水预测模型，提高了对英国职位的薪水预测准确度。刘睿伦、叶文豪、高瑞卿等人图片使用 ICD-CLAS 分词软件和 TF-IDF 法计算关键词权值，创建词条矩阵，使用 k-均值归类法，分析了网络招聘信息。韦家铎图片使用了三层 BP 神经网络，采用小批量梯度下降法作为收敛方向搜索算法，结合了附加动量法和自适应学习速率法两种方法的混合算法 Nadm 来对模型进行优化，加快了模型训练的收敛速度，并进一步提高了预测薪资结果的准确率。

1.3 研究内容与方法

本文的主要工作是研究机器学习技术在薪资预测方面的应用。通过机器学习技术对不同综合指标及各因素进行分析为数据分析岗位进行薪资预测，为用户提供更准确的建议和信息，使求职者能够从海量的职位信息中选择更合适的职位，提升用户体验图片。

本文主要的研究内容和使用的方法如下：

（1）查阅相关文献，基于机器学习深入研究数据分析行业的薪资预测问题；

（2）运用爬虫技术，在各个权威统计网站上爬取到了大量的有关数据分析岗位薪资或间接与其有关的近期数据；

（3）运用随机缺失森林算法来填补数据缺失值使其完整；

（4）利用聚类检测离群点去除不合实际的无效值；

（5）用随机森林在庞大的数据库中选出价值量最高，最重要的数据；

（6）利用人工神经网络进行学习，并构造出理想的模型。

第二章、相关理论概述

2.1 网络爬虫介绍

目前，通过使用 Shell Script、Python、R、C++ 等均可实现网络爬虫，并且相关支持库与示例代码丰富，是采集网络数据的较优方法。相较人工录入网上信息，爬虫拥有以下优点：快速便捷、准确度高、可重复能力强。

网络爬虫又称为网络蜘蛛，分为通用爬虫和聚焦爬虫。通用爬虫的基本原理如下：

1、将待爬取的网页链接地址即种子放入待爬取队列中；

2、从队列中取出 URL，进行页面请求及页面数据提取；

3、将下载到的页面保存到内存，放入特定工具中进行解析；

4、运行下一次数据爬取，直到爬取完所有可供爬取的数据或满足截止爬取条件为止。

2.2 随机森林模型介绍

决策树算法是一种逼近离散函数值的方法，一般可用于分类或者回归——对离散值的分类、对连续值的回归。决策树算法作为一种典型的分类方法，通过构造树状结构的决策流程，来发现数据中蕴涵的分类规则。如何构造精度高、规模小的决策树是决策树算法的核心内容。

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。一般情况下，训练样本数据集是根据实际需要有历史的、有一定综合程度的，用于数据分析处理的数据集；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、校正和修下的过程，主要是用新的样本数据集（称为测试数据集）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，提高决策树模型的泛化能力。

决策树模型具有易于理解，训练对数据量的要求不高，能够处理数值型数据和分类数据，能够完成多分类任务，具有高结果解释性，模型训练结果易于验证等优点，但同时决策树模型也具有容易过拟合、模型鲁棒性不足、模型训练复杂度过高（本质上是一个NP难问题）、对数据的平衡性过于敏感等缺点。

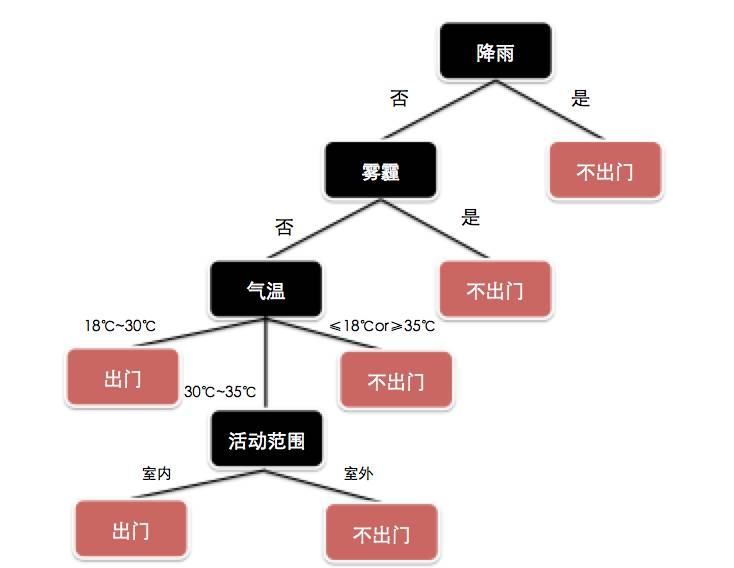


图 2.2-1 决策树算法预测过程

随机森林是一个包含多个决策树的分类器，可以用来进行分类和回归任务，并且其输出的类别是由个别树输出的类别的众数而定。其实从直观角度来解释，每棵决策树都是一个分类器（假设现在针对的是分类问题），那么对于一个输入样本，N 棵树会有 N 个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别指定为最终的输出，这就是一种最简单的 Bagging 思想。

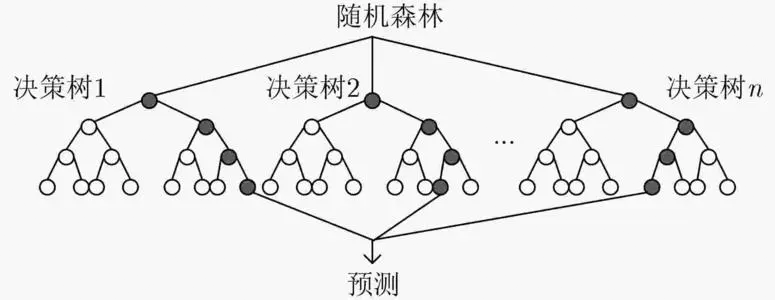


图 2.2-2 随机森林与决策树

随机森林训练算法把 Bagging 算法的一般技术应用到树学习中。随机森林本质上属于 Bagging 算法的应用，属于集成算法的一类。它与集成算法、Boosting 算法和 Bagging 算法的关系如下图：

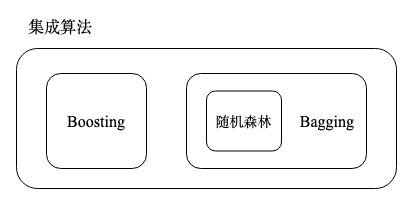


图 2.2-3 随机森林与集成算法、Boosting 算法和 Bagging 算法关系

在训练结束之后，对未知样本 x 的预测可以通过对 x 上所有单个回归树的预测求平均来实现：

图片

图 2.2-4 随机森林的预测

随机森林是由很多决策树构成的，不同决策树之间没有关联。当我们进行分类任务时，新的输入样本进入，就让森林中的每一棵决策树分别进行判断和分类，每个决策树会得到一个自己的分类结果，决策树的分类结果中哪一个分类最多，那么随机森林就会把这个结果当做最终的结果。随机森林的应用过程如下图：



图 2.2-5 随机森林的应用过程

随机森林是一种灵活的、便于使用的机器学习算法，即使没有超参数调整，大多数情况下也会带来好的结果，与同属委员会机器学习算法的 Bagging 算法和 Boosting 算法相比较，Bagging 可以从某个角度看是随机森林算法的一个特例 —— 在构造每一棵分类树的时候，如果允许选择进行分裂尝试的变量为全变量，则随机森林算法将退化成 Bagging 算法，但同时随机森林算法由于每次可以规定只随机选择少量的变量进行分裂尝试，因此可能有一些与结果高度相关的变量并不能每一次均被纳入待选变量中，从而有效避免了 Bagging 算法的每一个 bootstrap sample 构造的分类树结构过于相似的“死板性”。而 Boosting 算法同样和随机森林关系密切 —— 两者都通过 bootsrap 自助法选取样本，都是通过不断叠加简单弱分类器的优势形成的具有较高判别精度的分类器，两个分类器都不会产生过拟合效果，但相比较之下，由于 Boosting 算法属于迭代算法，每一次生成的弱分类器都依赖于上一个产生的分类器，而随机森林算法则每一棵生成的树都是独立的，因而在算法具体实现上，随机森林可以通过并行计算加速结果计算，而 Boosting 算法无法进行并行运算。

2.3 神经网络模型介绍

BP（Backpropagation，反向传播）神经网络是 1986 年由 Rumelhart 和 McClelland 为首的科学家提出的概念，特点是在训练模型的过程中采用了误差反向传播的方法，是目前主流的神经网络模型。BP 神经网络的拓扑结构包括一个输人层，一个或多个隐藏层和一个输出层。每层中都包含着若干个神经元。

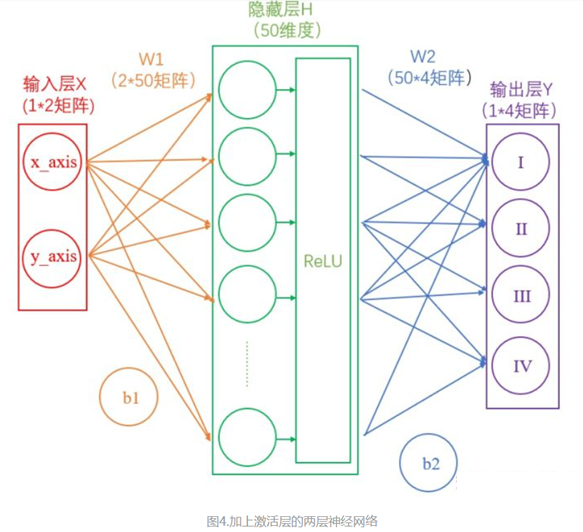


图 2.3-1 BP神经网络示意图

BP 算法主要由两个阶段组成：激励传播与权重更新。第一阶段中每次迭代的传播环节包含两步：前向传播与反向传播。前向传播阶段会将输入层数据送入网络以获得激励响应，而反向传播阶段会将激励响应同训练输入对应的目标输出求差，从而获得输出层和隐藏层的响应误差。反向传播阶段，对于每个突触上的权重，会按照以下步骤进行更新：1. 将输入激励和响应误差相乘，从而获得权重的梯度；2. 将这个梯度乘上一个比例并取反后加到权重上。这个比例（百分比）将会影响到训练过程的速度和效果，因此我们称之为“训练因子”。梯度的方向指明了误差扩大的方向，因此在更新权重的时候需要对其取反，从而减小权重引起的误差。

前向传播与反向传播反复循环迭代，直到网络对输入的响应达到满意的预定的目标范围为止，模型即停止训练。

反向传播要求有对每个输入值想得到的已知输出，来计算损失函数梯度。因此，它通常被认为是一种监督式学习方法，虽然它也用在一些无监督网络（如自动编码器）中。它是多层前馈网络的 Delta 规则的推广,可以用链式法则对每层迭代计算梯度。反向传播要求人工神经元（或“节点”）的激励函数可微。

第三章、数据爬取与预处理

针对本次研究主题，我们选取了两个数据方向，分别是职位信息与指数信息。职位信息有助于让我们得出薪资与特定地区、企业、学历和经验等关系，指数信息有助于让我们进一步研究城市对薪资的影响。

3.1 数据分析岗定义与薪资数据的选取和获取

3.1.1 数据分析岗的定义

职位是随组织结构定的，而岗位是随事定的，也就是我们常说的因事设岗。岗位是组织要求个体完成的一项或多项责任以及为此赋予个体的权力的总和。一份职位一般是将某些任务、职责和责任组为一体；而一个岗位则是指由一个人所从事的工作。而此次岗位选取主要以考虑在发达城市中的职位为对象。

数据行业的每一个环节都需要专业人员来完成。因此，需要培养和造就一支掌握数据技术、懂管理、有数据应用经验的数据建设专业团队。当前数据相关人才的紧缺严重阻碍了数据市场的发展。据 Gartner 预测，到 2017 年，全球将新增 440 万个数据相关工作岗位，25% 的组织将设立首席数据官职位。数据分析相关岗位需要能够综合驾驭数学、统计学、数据分析、机器学习、自然语言处理等多方面知识的复合型人才。未来，数据分析方面的人才缺口将达到 100 万左右。在各个行业，数据分析领域的高端人才将成为热门人才，包括大数据数据开发工程师、数据分析师、数据架构师、数据后端开发工程师、算法工程师等多个方向。

3.1.2 数据分析岗数据的选取

本文在选取数据时考虑了招聘网站权威性、数据代表性、时间段等因素。招聘信息的采集方式一般可分为线上与线下，分别对应企业在线发布与线下招聘会、人才市场公告。介于采集便利程度与数据量大小，本文所需的数据分析方向职位岗位的相关招聘数据选取了采集线上数据。目前主流的招聘信息发布网有前程无忧网、拉勾网、实习僧等，本文选取前程无忧平台作为数据来源，采用网络爬虫技术获取相关数据。前程无忧成立于 1999 年，是一家网络招聘服务提供商，网站目标有两大部分：致力于为积极进取的白领阶层和专业人士提供更好的职业发展机会。同时，网站致力于为企业搜寻、招募到最优秀的人才。该平台提供报纸招聘、网络招聘、招聘猎头、培训测评和人事外包在内的人力资源服务，最初全国包括香港在内的 26 个城市设有服务机构，至今已经增加至 104 个城市。该平台在数据分析方向岗位需求更新量大，时效性强，覆盖地域广，权威性高。

在数据获取的时间段方面，综合考虑到我国毕业时间段与暑期时间段，本文选取了 2021 年 5 月 20 日至 2021 年 7 月 20 日这一大量劳动力涌入劳动市场的时间段作为数据收集时间段。综合考虑数据抓取难度、数据量、网站权威性、数据代表性、程序便利性与数据处理便利性等因素，本文选择使用 Python 对前程无忧网站数据分析师岗位约 1 万条记录进行爬取。该数据涵盖了北京市、上海市、重庆市、天津市、西安市、成都市、昆明市、长春市、合肥市、郑州市、南京市、苏州市、哈尔滨市、广州市、深圳市、东莞市、长沙市、福州市、武汉市、济南市、青岛市、杭州市、宁波市、大连市和沈阳市共 25 个全国热门城市作为分析对象，收集的数据包括岗位标题、学历要求、工作经验要求、职位发布日期、职位标签、职位描述、企业名称、企业类别、工作地点和薪水数值。

3.1.3 数据分析岗薪资数据的获取

爬虫系统模块主要分为HTTP 请求模块（requests 包）和HTML 解析模块（bs4 包与 lxml 包）。根据前途无忧网站架构，本文数据起先使用了 selenium web driver + chromium 的方式爬取，并使用 web driver 的 find\_elements\_by\_css\_selector 方法进行元素选取。该方法耗时久、效率低，最终被弃用。后来选择使用 requests 包进行数据爬取，获取到 HTML 源代码再使用 bs4 包进行解析。过程中，前程无忧页面使用了反爬技术，PC 端网页使用了懒加载（lazy loading）策略反爬。最终本文选定使用 requests 包爬取移动端网页，并使用 bs4 包对获取到的 HTML 进行解析，获取到搜索页面下的标题、标签、薪资、企业名称、企业类型、工作地点、学历要求和工作经验要求，并提取得到岗位详情信息对应的 URL 地址，再进一步对该页面发起请求，提取得到描述信息和岗位发布日期。

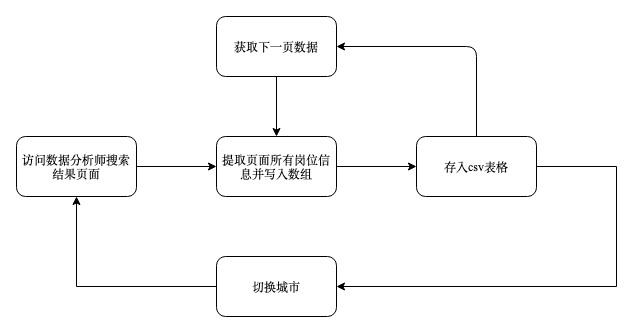


图 3.1-1 岗位基础信息爬取流程



图 3.1-2 岗位描述信息与发布日期爬取流程

为加快访问速度，本文在设计时分离了网络请求模块与数据解析模块，使用了 threading 模块分线程运行数据请求与解析函数，显著提高了运行速度与执行效率。

3.2 热门城市基本指数的获取与挑选

3.2.1 城市基本指数

城市基本指数，是指当前存在反映中国各省份地区城市一定社会经济发展方面的指标及其数值，反映我国城市经济社会发展程度。目前，我国各大城市在工业化、城市化、农业现代化方面取得显著突破，如何评估以及核算一个城市发展的速度、发展的质量、取得的优势之处、所存在的不足之处，关键便在于对城市基本指数的统计与分析。

十一五规划《纲要》描述了九个方面的经济社会发展指标，全篇定量的指标一共有 39 个，其中最主要的指标一共有 22 个指标，反映了经济、经济结构、人口、资源、环境、公共服务和人民生活方面的发展程度。这在城市当中，可以反映城市综合实力、农业、民政、民生、科技金融、教育文化、交通邮电、环境卫生、建筑房产工业、对外经济和旅游方面的发展程度。

城市基本指数是一个地区综合经济实力的体现，是考核各级地方政府的重要指标，由于涉及指标数量多，牵扯行业面广，如何将核算用的大量指标凝聚在一起，再进一步如何将各指标进行综合分析，是在新的核算体系下需要解决的问题。本文就基于城市基本指数的定义对数据进行分析。

3.2.2 热门城市基本指数的挑选与获取

在指标的挑选方面，我们根据时效性和代表性两个指标，综合考虑相关因素，筛选出我们需要选取的热门城市基本指数。我们着重考虑了指数是否具有代表性这一指标，如综合项目中的 GDP，GDP 增长值以及第一、二、三产业，对此类指数数据的收集和分析可以较准确的研究当今背景下对就业情况进行相对准确的预测。在中国指数各版块下，我们针对初步判断，筛选留下了 51 个指标供爬取。

对于热门城市的参数爬取，进行了人工参数采集，将参数与城市分别放入列表，且顺序一一对应，以保证城市参数匹配准确性。另外，对于城市选择，选取了相对热门或是发达的城市，对岗位薪资研究有更好参照性。然后选择使用 requests 包对知网数据库进行数据爬取，获取到 HTML 源代码再使用 lxml 包进行解析，最终获得了热门城市51个基本指数的相关原始数据。

3.3 数据分析岗薪资数据预处理

3.3.1 岗位数据预处理

数据预处理是数据挖掘过程中的一个重要步骤。数据的质量，直接决定了模型的预测和泛化能力的好坏。它涉及很多因素，包括：准确性、完整性、一致性、时效性、可信性和解释性。而在真实数据中，我们拿到的数据可能包含了大量的缺失值，可能包含大量的噪音，也可能因为人工录入错误导致有异常点存在，非常不利于算法模型的训练。数据清洗的结果是对各种脏数据进行对应方式的处理，得到标准的、干净的、连续的数据，提供给数据统计、数据挖掘等使用。数据预处理的主要步骤分为：数据清理、数据集成、数据规约和数据变换。

本文研究的实验数据主要是以表格为存储形式的薪水职位数据，对于文本和数字数据进行分析的时候，需要对于文本和数字进行预处理，对于之后我们提出的薪水预测模型的效果有较大的影响。

3.3.2 岗位数据清洗

在完成数据收集工作后，我们使用 pandas 的 read\_csv 方法读取了上文爬取后输出的数据包 csv 表格文件，并将其转换为 DataFrame，执行了 pd.head 方法，输出如下清洗前岗位数据样例：

表 3.3-1 清洗前岗位数据样例



对于该数据，我们进行了如下步骤的数据清理工作，务求在后续神经网络的构建中，数据能够更有代表性、更真实准确且更容易使用。

（1）原始数据没有列名，我们为其添加列名，分别是 `URL`、`Title`、`Zone`、`Salary`、`Company Type`、`Degree and Experience`、`Company Name`、`Publish Date`、`Description`。

（2）我们对获取的职位薪水、地区数据进行缺失值处理。通常数据缺失的处理方法有：忽略数据、人工填补缺失值、常量填补缺失值、均值填补缺失值。

我们从中选取了最为可靠通用的方案，即忽略数据，原因如下：

1. 我们的数据量很大，数据量不够时候我们可以继续爬取；
2. 人工、常量、平均值会带来额外的噪声；

（3） 地区数据难以修补。原数据存在只精确到“省”的地区，而我们要预测到热门地级市，省难以预测到市；

综上所述，我们对于职位薪水、地区数据选择直接忽略缺失数据的做法，避免了额外的误差。

（3）我们需要对薪水数据做去重处理。介于网络爬虫与工作地区的重复性，存在不少相同工作被爬取录入到表格之中。我们通过选取爬取得到的 URL 和公司名、地点、职位名和学历经验要求，对重复的数据进行去重处理。

（4）我们需要对地区进行标准化处理。目前地区的存储分为“省”、“市”、“市-区”三种形式。我们选择使用 split 函数截取前一部分，并且与标准地名进行判断，从而对不合规的数据进行删除处理。

（5）我们需要对薪资数据进行标准化处理。目前薪资单位分为“万”、“千”/“天”、“月”、“年”，我们选择将所有数据转换为“个”/“月”。同时描述信息和标题中存在“13 薪”、“年底双薪”这样的字段。我们选择将其提出，用以优化我们的薪资数据。同时，我们对区间和薪资数区间进行提取，在表格中创建了三个新列（`Max Salary`、`Min Salary`、`Avg Salary`），用以进一步处理薪资数据。

（6）我们需要对学历和经验数据进行提取。原文中存储的格式为“学历”、“经验”、“学历-经验”。我们选择用前程无忧网搜索页面筛选框的的学历和经验数据进行提取，分离学历和经验数据，并存储到新列中（`Experience`、`Degree`）。

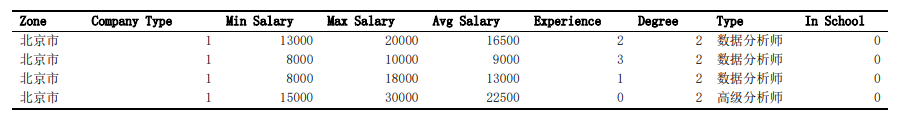
（7）我们对标题进行提取，根据 df.columns['Title'].unique() 所给出的值，我们划分为数据工程师、高级分析师、数据分析师三类别。

（8）我们对清洗出结果的表格去除了已经解决了的原始列，即去除了 Title、Salary、Degree And Experience、Company、Publish Date、Description 共六列。

（9）我们将上述结果得到的 DataFrame 的 Experience、Degree、Company Type 的取值转化为整型，便于后续学习。

清洗完成后，样例数据如下表：

表 3.3-2 清洗后岗位数据样例



3.4 热门城市综合指标的预处理

3.4.1 缺失值处理

由于在一些城市的部分年份并没有相关机构发布数据，因此爬取得到的不同综合指标数据存在不同程度上的缺失。考虑到指数的时效性与准确性，我们没有对指标进行人工填充或均值填充，而是选择弃用这一年的指标，使用 2019 年的数据，或对于数据过度稀疏、缺失值过多的直接弃用这一分类数据。

表 3.4-1 可以舍弃的数据例子（右表：女性人口数），需要保留的数据例子（左表：GDP 数据）

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 年份 | 数据个数 | 城市总数 | 是否存在缺失 |  | 年份 | 数据个数 | 城市总数 | 是否存在缺失 |
| 2020 | 25 | 25 | 否 |  | 2020 | 11 | 25 | 是 |
| 2019 | 25 | 25 | 否 |  | 2019 | 14 | 25 | 是 |
| 2018 | 25 | 25 | 否 |  | 2018 | 14 | 25 | 是 |
| 2017 | 25 | 25 | 否 |  | 2017 | 12 | 25 | 是 |
| 2016 | 25 | 25 | 否 |  | 2016 | 9 | 25 | 是 |
| 2015 | 0 | 25 | 是 |  | 2015 | 10 | 25 | 是 |

3.4.2 异常值处理

在数据收集过程中，我们遇到了诸如 2020 年广州市公共图书馆数反而超过广东省数，涨幅达到数十倍甚至百倍，或是指标上涨幅度 700% 的异常情况。针对这一情况，我们在图书馆数据库中查阅了相关年鉴，使用出版物中数据替换错误数据。

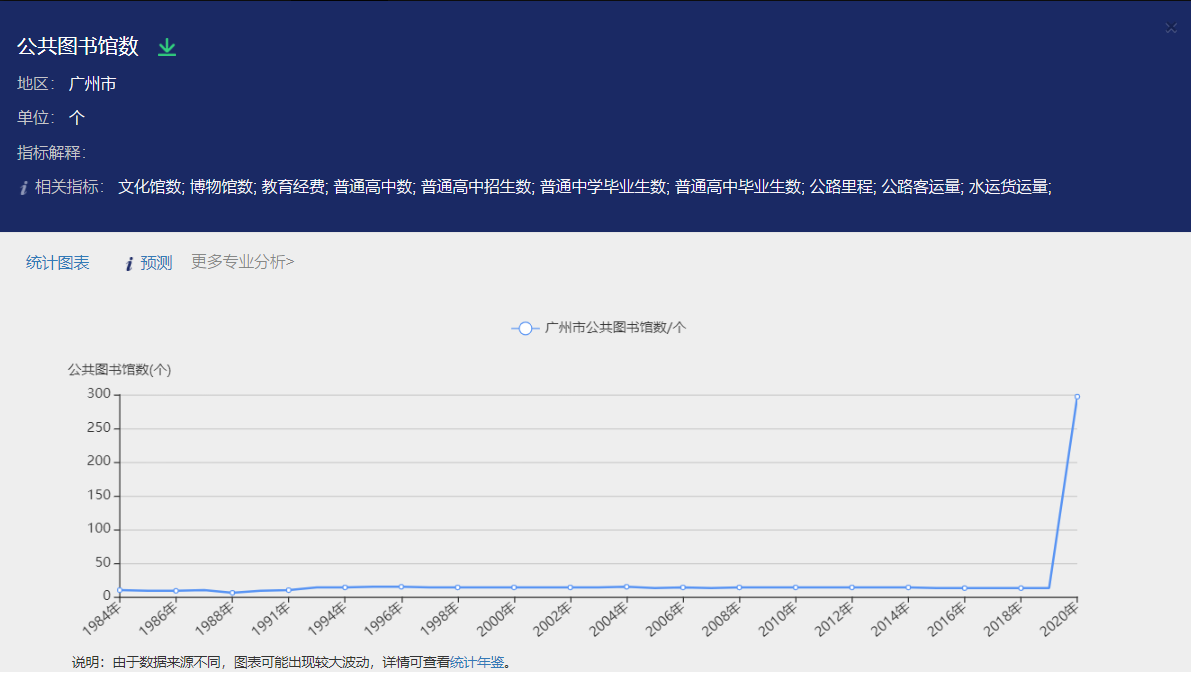


图 3.4-1 异常值案例 —— 广州市公共图书馆数

3.4.3 特征筛选

挑选数据并完成整理后，我们对欠拟合的数据进行筛选，如高技术产业主营业务收入（以省为单位），尽管此项数据能很好反应各省的高技术产业集中情况和业务收入，但出于各网页收集到的数据界限不一的原因，因此也没有到其他网址进行数据爬取来填充数据。基于数据的完整程度（是否有缺失，是否准确）、与薪资高低关联程度，筛选出 25 个数据，下表是我们最终选取的热门城市基本指数：

表 3.2-1 选取的热门城市基本指数

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 指标类型 | 单位 | 年份 |  | 指标类型 | 单位 | 年份 |
| GDP 数据 | 亿元 | 2020 | 房屋竣工面积数据 | 万平方米 | 2019 |
| GDP 增长率数据 | % | 2020 | 失业保险参保人数数据 | 万人 | 2020 |
| GDP 指数数据 | (上年=100) | 2019 | 公共图书馆藏书量数据 | 万册 | 2019 |
| 城镇登记失业人员数据 | 万人 | 2019 | 普通高等学校数数据 | 所 | 2019 |
| 第三产业增加值数据 | 亿元 | 2020 | 人口自然增长率数据 | % | 2019 |
| 第二产业增加值数据 | 亿元 | 2020 | 卫生机构数数据 | 个 | 2019 |
| 出生人口数数据 | 人 | 2019 | 医院数数据 | 个 | 2019 |
| 房屋建筑施工面积数据 | 万平方米 | 2019 | 总户数数据 | 户 | 2019 |
| 房地产开发住宅投资额 | 万元 | 2019 | 小学学校数数据 | 所 | 2019 |
| 第一产业增加值数据 | 亿元 | 2020 | 总人口数数据 | 万人 | 2019 |
| 房地产开发投资额数据 | 万元 | 2019 | 实际利用外资额数据 | 万美元 | 2019 |
| 商品房销售面积数据 | 万平方米 | 2019 | 外商直接投资额数据 | 万美元 | 2019 |
| 社会消费品零售总额 | 万元 | 2019 |  | | |

3.4.4 小结

续表 3.2-1

接上页

我们在中国经济社会大数据研究平台上共爬取51个与数据分析岗相关的指标，为了保证数据的质量，我们在保留的数据中选择近六年中相较完整同时具有时效性的一年作为分析的数据。结果如下：

表 3.4-2 指数汇总表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 指标类型 | year | 北京市 | 上海市 | 重庆市 | ... | 宁波市 | 大连市 |
| GDP数据 | 2020 | 36102.6 | 38700.58 | 25002.79 | ... | 12408.7 | 7030.4 |
| GDP增长率数据 | 2020 | 1.2 | 1.7 | 3.9 | ... | 3.3 | 0.9 |
| GDP指数数据 | 2019 | 106.1 | 106 | 106.3 | ... | 106.8 | 106.5 |
| 人口自然增长率数据 | 2019 | 7.44 | -2.3 | 2.79 | ... | 2.13 | 0.91 |
| ...... | ..... | ...... | ...... | ...... | ... | ...... | ...... |
| 总人口数数据 | 2019 | 2154 | 2428 | 3124 | ... | 608.4707 | 598.7 |
| 实际利用外资额数据 | 2019 | 1421000 | 1904800 | 236500 | ... | 236341 | 87000 |
| 外商直接投资额数据 | 2019 | 1421299 | 1904791 | 1031042 | ... | 236341 | 64617 |

3.5 数据分析岗基本特征分析

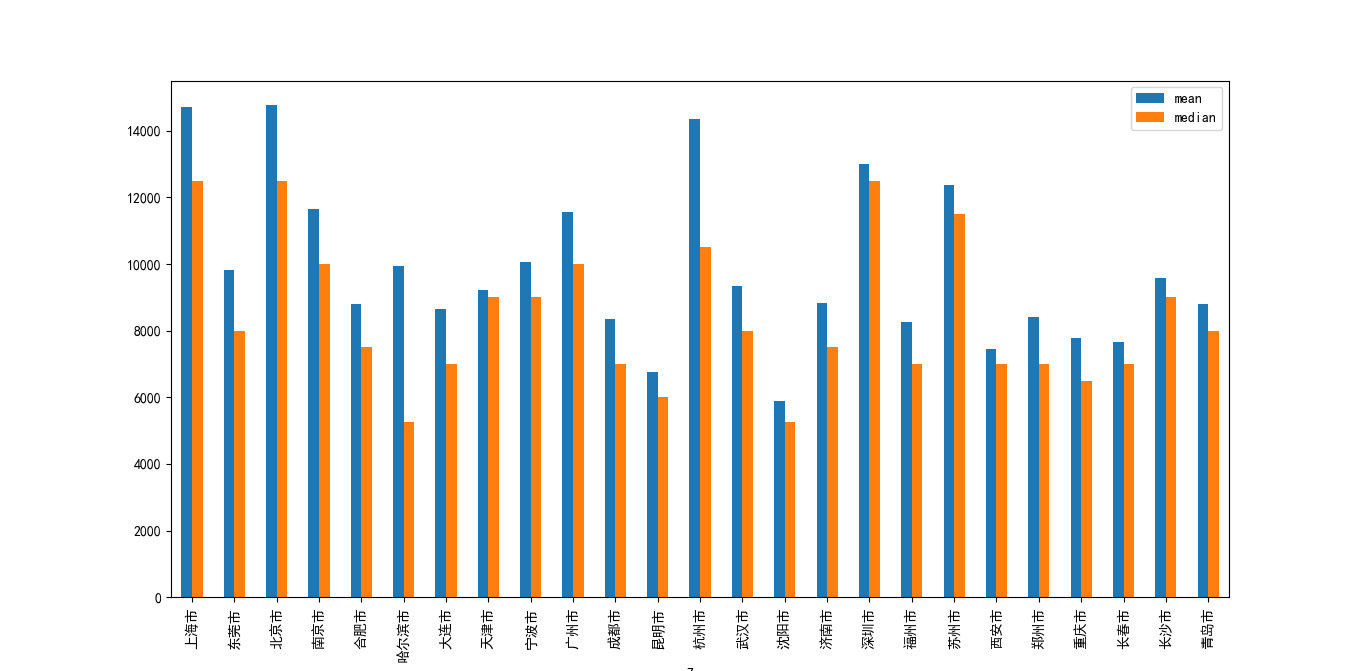


图 3.5-1 各个城市数据分析师的薪资均值与中位数

通过上图，可以发现各个城市的均值与中位数差距不大，但是城市之间区别非常大；北上深杭这四个城市的工资水平远远高于其他城市，广州苏州稍为靠后；靠前的这几个城市也是全国经济最为发达的几个城市，而且都发布在东部，中部城市薪资水平普遍低一些。

由此可见，即便是相同类型的职业，但在不同城市有着不一样的薪资，而且各个城市在数据分析师的薪资均值与中位数差值较大，证明不同发展水平的城市对数据分析师这类技术要求较高的职业有所不同，代表该职业人员的能力有所高低，而薪资恰与此紧密相关，该城市的当地教育以及对外地的牵引力（吸引人才）的影响非常重要，而面对选择此类型职业时，我们应该了解自己心目中的薪资与自己能力是否相符合，以致于能够选择一座适合自己的城市生活。

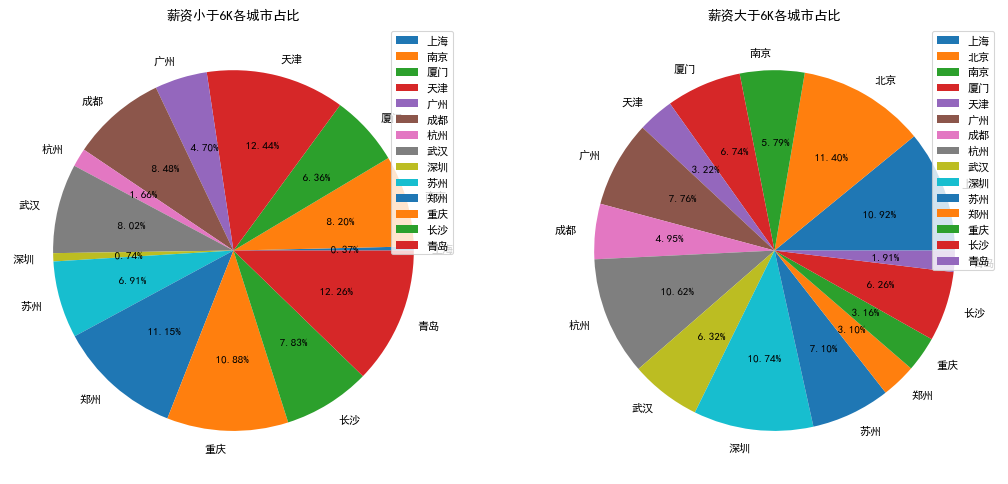


图 3.5-2 不同城市数据分析师薪资饼状图

以上饼状图中可看出，各城市间内数据分析师岗位的薪资水平差异非常大。薪资水平小于6k的岗位主要集中分布于重庆、郑州、成都、武汉等中西部城市；而北上广深杭等城市基本不存在薪资水平小于六千的情况；而薪资水平大于 6 千的岗位分布区别则非常明显，北上广深杭的所占比重直接升至 10% 以上，重庆、郑州、成都、武汉等城市所占比重则少到 3% 左右。

对于图 3.5-2，这幅饼图更加直观的看到带高底薪的数据分析师所集中的城市占比，低薪岗位主要分布于中西部城市，而高薪主要分布在东部沿海的发达城市。这当中包含有政策，经济，企业，交通，外贸，高技术人才等众多因素，都影响城市发展水平，从而影响对数据分析师有相对较高的要求，因此薪资也相对较高，由此看出，我们选择除了看城市经济水平，还要注意其他因素对岗位的影响。

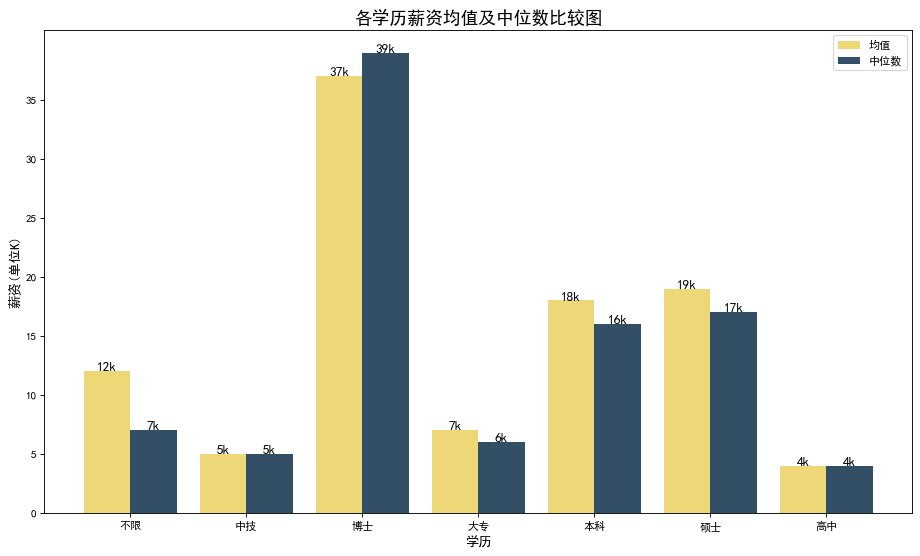


图 3.5-3 各学历薪资均值及中位数比较

从市场需求来看，要求本科占据绝大多数；该岗位对学历的门槛并不至于很高，要求硕士的较少，而博士则是凤毛麟角；不过这也从侧面反映了社会上学历的分布状况，本科生及大专生占绝对多数；虽然市场需求量大，但是本科生就业压力也很大。

由此得出数据分析师这一职业，相对于社会上其他职业，算是中上水平的职业，所以我们若想谋求到数据分析师的职业，念到本科生即可，对于硕士、博士，若有对薪资有较高要求且水平够高，可以考虑。

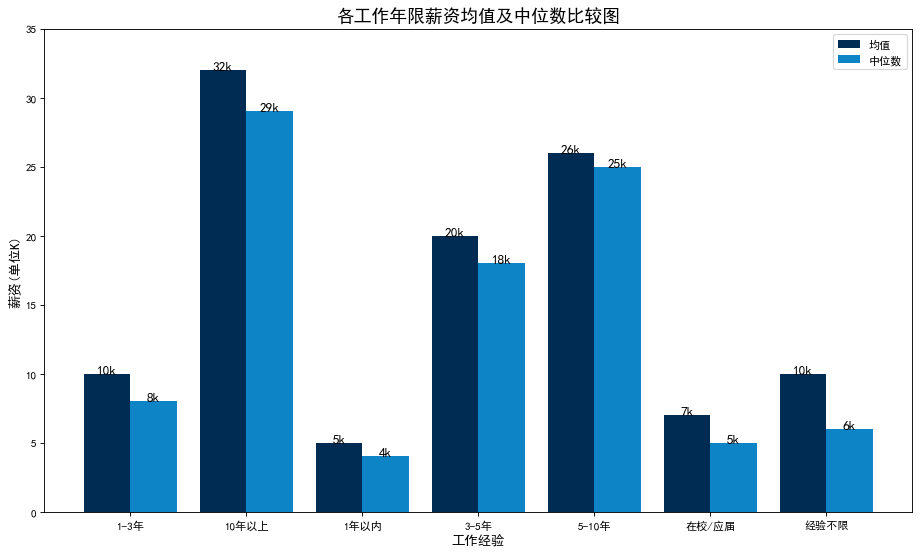


图 3.5-4 工作时间(经验)对应薪资图

应届生的工资较低和工作经验一年内的差距并不大。但是随着工作年限的增加，迈过 3 年工作经验这个门槛，工资水平将直线上升。者说明数据分析岗位是一个需要累积丰富经验的岗位。

第四章、基于神经网络的数据分析岗薪资预测

4.1 基于随机森林算法的特征选择

4.1.1 基于随机缺失森林的热门城市综合指标填充

在获得相较为完整的数据后，我们要对部分缺失的热门城市的综合指标进行填充，随机森林算法本身就具有很好的分类精度，从而也更进一步确保了得到的填补值的准确性和可靠性。我们利用 sklearn 的 RandomForestRegressor 即随机森林回归方法处理我们的数据。

数据共二十五列，我们将拥有完整数据的14列提取出来作为已知特征矩阵，循环其余11列，然后划分 Xtrain 和 Ytrain，Xtest 和 Ytest。Xtrain 和 Ytrain 是缺失值所在列除去缺失值所在的行，Xtest 是缺失值所在的行除去缺失值所在的列，Ytest 其实就是缺失值，最后我们要用 predict 得到的值来填补。

原数据如下表：

表 4.1-1 原始数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 北京市 | 24 | non-null | float64 | 广州市 | 25 | non-null | float64 |
| 上海市 | 24 | non-null | float64 | 深圳市 | 25 | non-null | float64 |
| 重庆市 | 25 | non-null | float64 | 东莞市 | 22 | non-null | float64 |
| 天津市 | 24 | non-null | float64 | 长沙市 | 23 | non-null | float64 |
| 西安市 | 24 | non-null | float64 | 福州市 | 25 | non-null | float64 |
| 成都市 | 24 | non-null | float64 | 武汉市 | 23 | non-null | float64 |
| 昆明市 | 23 | non-null | float64 | 济南市 | 24 | non-null | float64 |
| 长春市 | 24 | non-null | float64 | 青岛市 | 22 | non-null | float64 |
| 合肥市 | 25 | non-null | float64 | 杭州市 | 25 | non-null | float64 |
| 郑州市 | 24 | non-null | float64 | 宁波市 | 24 | non-null | float64 |
| 南京市 | 25 | non-null | float64 | 大连市 | 25 | non-null | float64 |
| 苏州市 | 25 | non-null | float64 | 沈阳市 | 25 | non-null | float64 |
| 哈尔滨市 | 22 | non-null | float64 |  |  |  |  |

可以看出，原数据存在较多行不足25个非空数据的情况，用得到的模型进行未知特征值预测，处理所得 DataFrame 如下：

表 4.1-2 预测后数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 北京市 | 25 | non-null | float64 | 广州市 | 25 | non-null | float64 |
| 上海市 | 25 | non-null | float64 | 深圳市 | 25 | non-null | float64 |
| 重庆市 | 25 | non-null | float64 | 东莞市 | 25 | non-null | float64 |
| 天津市 | 25 | non-null | float64 | 长沙市 | 25 | non-null | float64 |
| 西安市 | 25 | non-null | float64 | 福州市 | 25 | non-null | float64 |
| 成都市 | 25 | non-null | float64 | 武汉市 | 25 | non-null | float64 |
| 昆明市 | 25 | non-null | float64 | 济南市 | 25 | non-null | float64 |
| 长春市 | 25 | non-null | float64 | 青岛市 | 25 | non-null | float64 |
| 合肥市 | 25 | non-null | float64 | 杭州市 | 25 | non-null | float64 |
| 郑州市 | 25 | non-null | float64 | 宁波市 | 25 | non-null | float64 |
| 南京市 | 25 | non-null | float64 | 大连市 | 25 | non-null | float64 |
| 苏州市 | 25 | non-null | float64 | 沈阳市 | 25 | non-null | float64 |
| 哈尔滨市 | 25 | non-null | float64 |  |  |  |  |

4.1.2 招聘数据集的规范化

数据标准化（归一化）处理是数据挖掘的一项基础工作，不同评价指标往往具有不同的量纲和量纲单位，这样的情况会影响到数据分析的结果，为了消除指标之间的量纲影响，需要进行数据标准化处理，以解决数据指标之间的可比性。原始数据经过数据标准化处理后，各指标处于同一数量级，适合进行综合对比评价。

在通过随机缺失森林的数据填补后，需要消除每一个指标的量纲，这里运用了极差标准化，公式如下：

图片

得到无量纲数据如下表：

表 4.1-3 无量纲数据

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 北京市 | 上海市 | … | 大连市 | 沈阳市 |  | mean | max | min | range |
| GDP数据 | 0.621264 | 0.698777 | … | -0.24613 | -0.25982 |  | 15279.84 | 38700.58 | 5183.8 | 33516.78 |
| GDP增长率数据 | -0.15394 | -0.10343 | … | -0.18424 | -0.19434 |  | 2.724 | 5.2 | -4.7 | 9.9 |
| GDP指数数据 | -0.12739 | -0.15442 | … | -0.01929 | -0.64091 |  | 106.5714 | 107.9 | 104.2 | 3.7 |
| … | … | … | … | … | … |  | … | … | … | … |
| 总人口数数据 | 0.36441 | 0.459782 | … | -0.17695 | -0.12241 |  | 1107.072 | 3124 | 251.0586 | 2872.941 |
| 实际利用外资额数据 | 0.480603 | 0.739092 | … | -0.23214 | -0.19046 |  | 521481.9 | 1904800 | 33154.72 | 1871645 |
| 外商直接投资额数据 | 0.457815 | 0.71612 | … | -0.26699 | -0.21333 |  | 564364.9 | 1904791 | 33000 | 1871791 |

对于数值化处理后得到的数据分析岗薪资数据我们同样需要使用极差标准化方法对数据进行规范化处理，得到结果样例如下表所示：

表 4.1-4 规范化处理后数据

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Zone | Company Type | Min Salary | Max Salary | Avg Salary | Experience | In School | Degree | Type |
| 北京市 | 0.1479 | 0.0216 | 0.022 | 0.0219 | 0.2152 | -0.0719 | 0.0921 | -0.1278 |
| 北京市 | 0.1479 | -0.003 | -0.011 | -0.008 | 0.4652 | -0.0719 | 0.0921 | -0.1278 |
| 北京市 | 0.1479 | -0.003 | 0.0153 | 0.0078 | -0.035 | -0.0719 | 0.0921 | -0.1278 |
| 北京市 | 0.1479 | 0.0317 | 0.0555 | 0.046 | -0.285 | -0.0719 | 0.0921 | 0.37223 |
| … | … | … | … | … | … | … | … | … |
| 济南市 | 0.1479 | -0.029 | -0.03 | -0.029 | -0.285 | -0.0719 | -0.408 | -0.1278 |
| 长沙市 | 0.1479 | -0.029 | -0.032 | -0.03 | -0.285 | -0.0719 | -0.408 | -0.1278 |
| 郑州市 | 0.1479 | -0.003 | -0.001 | -0.002 | -0.285 | -0.0719 | 0.0921 | -0.1278 |
| 福州市 | 0.1479 | -0.019 | -0.018 | -0.018 | -0.035 | -0.0719 | -0.158 | 0.87223 |

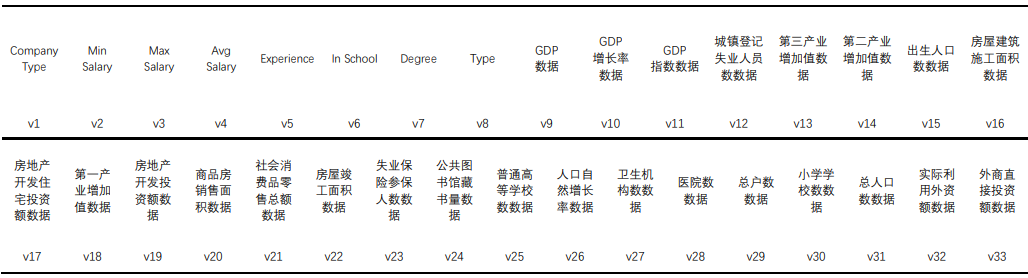
最后，我们需要将规范化得到的数据分析岗数据按岗位地区拼接上对应城市的规范化基本指数，并将代表地区的列 Zone 删除掉，从而形成我们最终用于模型训练的标准数据集。

4.1.3 基于随机森林的训练数据特征选择

在进行神经网络的预训练之前，我们需要先进行特征提取。我们选择了 sklearn 作为特征选择框架，导入 4.1.2 处理后的数据表格。

4.1.1 与 4.1.2 将所有数据标准化后，修改了列名，如下对应所示：

表 4.1-5 列名前后对应关系



如图所示，我们应该选择 v4 即平均薪资作为讨论对象，其他除了变量v2、v3外的 30 列数据作为特征，进行计算。我们使用了 RandomForestRegressor 即随机森林回归方法处理我们的数据。

我们访问了拟合属性 feature\_importances\_ 以获取特征重要性，它们被计算为每棵树内杂质减少的累积平均值和标准偏差。 随后我们对得到的结果使用 numpy 的 std 方法求标准差，并使用 yerr 属性传参，以显示选择后的特征重要性。

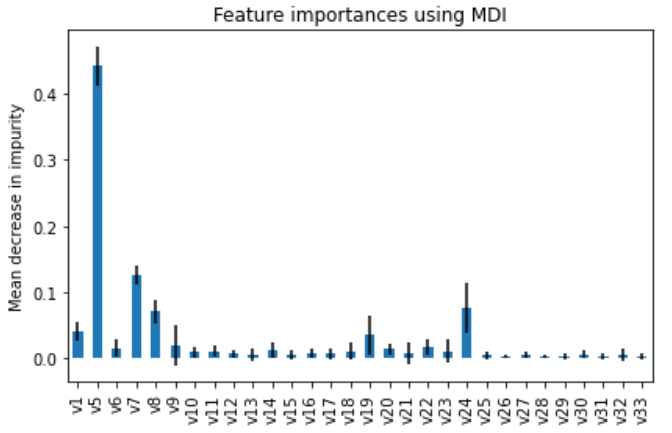


图 4.1-3 使用 MDI 得到的特征重要性。

我们将所计算得出的特征重要性，按照 0.05 为划分区间，大于 0.05 的我们可认为是重要特征。使用 sum 函数进行计算，满足这一条件的特征共有 24 列。

从学历方面来看，本科是一个分水岭；是否是本科，对薪资水平的影响还是很大的；而本科与硕士之间并没有特别大的差距；博士的薪资水平远远高于其他学历。

首先从均值与中位数来观察，样本数据大体差异并不大，除了图中“不限”学历的薪资属于特殊值以外，基本就业的薪资与学历挂上不可分割的关系了。从各个学历观察，我们不难得出，本科及以上学历能有较高薪资，这与随机森林进行变量选择的结果是相吻合的。

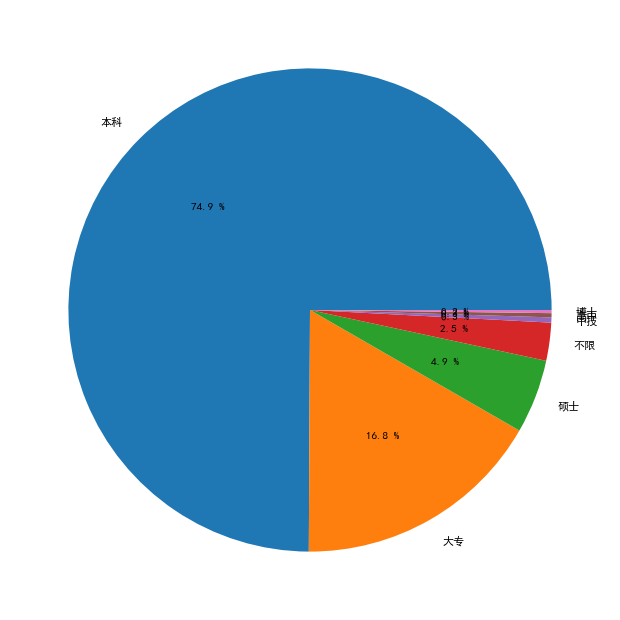


图 4.1-4 各个学历市场需求饼状图

由图 4.1-5 可知，市场上对于工作经验的要求还是很强烈的，1-5 年的经验要求占绝大多数。而应届生和 10 年以上的需求则非常少，工作经验是入职成功与否的关键因素，其次，可以看出数据分析师的工作内容并不需要非常的深度，只要有一定基础，再工作4到5年就可以达到数据分析信手拈来的效果。而这也点也代表我们在工作4-5年后可以进行转职或是创业以谋求更多财富或是阅历，职业的弹性限度也很大，同时由于数据分析所占领域之宽广，进行职业转型也并不止于非常困难。

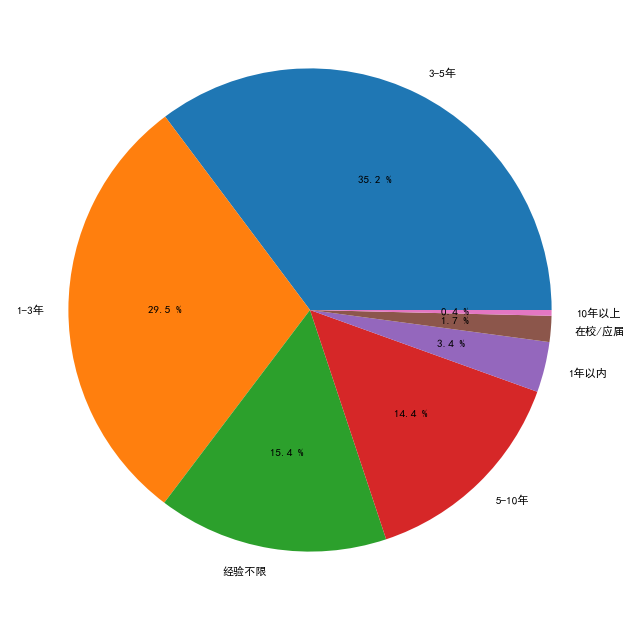


图 4.1-5 对员工经验要求饼状图

另外，对于数据分析这类职业而言，想要查找资料，图书馆的书籍和论文查找等功能服务必不可少，因此这也为此类岗位职工提供很大帮助。同样，图书馆的藏书等能够正向影响一个地区的人才实力，间接影响这个地区的平均薪资高低。

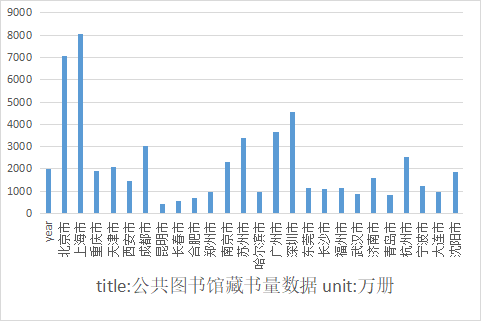


图 4.1-7 公共图书馆藏书量数据

通过上述分析，我们可以发现，基于随机森林对特征重要性的选择结果，与我们基于统计数据进行直接分析得到的结果相一致，这表明我们可以基于随机森林对特征进行提取的结果来进行变量选择，并完成后续神经网络模型的构建工作。

4.2 薪资预测模型的选择

4.2.1 回归模型

薪资预测是一种回归预测。机器学习中常用的回归预测模型有线性回归、多项式回归、岭回归、套索回归、弹性网络回归和神经网络。

线性回归是指由线性变量组成的回归模型，因此必须满足自变量和因变量之间的线性关系，即线性回归只适用于解决线性可分问题，线性回归的缺点是对异常值非常敏感。如果数据中存在大量的异常值，将会影响模型的精度。

多项式回归是线性回归的推广。指回归方程中存在指数大于 1 的自变量。当所有自变量的指数都等于 1 时，多项式回归方程退化为线性回归方程，因为自变量和因变量不是线性的，多项式回归可以解决非线性和复杂的问题，多项式回归模型的缺点是最佳指标难以确定，指标过高容易导致过拟合，指标过低容易导致欠拟合。

岭回归是在线性或多项式回归的数据特征中存在多重共线性（即自变量之间近似线性相关）时使用的一种优化技术。它通过在回归估计函数（损失函数）中加入一个平方偏移因子来减少多重共线性的影响。

套索回归与岭回归类似，是在线性或多项式回归的数据特征中存在多重共线性以及回归估计函数中存在偏移因子时所采用的一种优化技术，只有岭回归加上一个平方偏移，而套索回归增加了一个绝对值偏移。

弹性网回归是岭回归和套索回归的混合。在回归估计函数中加入平方和绝对偏移量，达到岭回归和套索回归的共同效果。

神经网络是应用最广泛的网络。理论上，三层 BP 网络可以逼近任意精度的连续函数，具有学习能力强、网络结构简单、计算量小等优点，但神经网络存在一些不足。神经网络存在的主要问题是网络拓扑结构、参数初始化和优化算法等超参数对模型的训练结果影响很大。如果超参数选取不当，很容易导致模型计算结果不理想。

4.2.2 模型比较

薪资水平众多因素相关，它不仅和个人的基础能力相关，也和不同的企业存在联系，是个人能力在企业中的具体体现。因此薪资水平与所有因素之间不是简单地线性关系，使用线性回归作为薪资预测的学习方法并不是很好的选择。多项式回归相比于线性回归具有更好的灵活性，能够解决非线性等复杂问题，但是对于最佳指数的选择要求比较严格，稍有偏差就容易发生过拟合和欠拟合。岭回归、套索回归和弹性网络回归都是对线性回归和多项式回归的优化，能够在一定程度上解决过拟合问题，相对于线性模型和多项式模型而言是更好的选择，但是它们仍无法解决线性回归和多项式回归中存在的问题，即线性回归无法解决非线性问题和多项式回归最佳指数难以选择。

相对而言，BP 神经网络能够逼近任意精度的连续函数，也就是说它无需考虑预测函数的指数问题，不存在多项式回归中选择指数的问题，同时也能够解决线性回归无法解决的非线性问题。从性能上讲，BP 神经网络具有强大的学习能力，即使是不同形式的样本数据，依旧能够逼近代表样本数据规律的函数;从可扩展性来说，BP 神经网络是通过自主学习来获取变量之间的规律，因此即使对数据集进行改动，依旧具有很好的可扩展性和适应性。

综上所述，由于 BP 神经网络具有较强的学习能力和良好的可扩展性，并且在数值预测方面也有较为广泛的应用。因此，薪资预测将优先选用三层 BP 神经网络模型来进行实现。

4.3 神经网络模型的预训练

4.3.1 训练集与测试集的划分

训练集与测试集的划分有三种方法，分别是留出法、交叉验证法和自助法。自助法在数据集较小，难以有效划分训练集和测试集的时候比较有效。自助法改变了原有数据的分布，这样会引入估计误差。在数据量充足的时候尽量考虑用留出法或者交叉验证法

综合考虑数据量、操作复杂程度与现有条件，我们选择使用留出法。我们运用随机抽取的方法划分训练集与测试集。在我们选择使用的神经网络包中的 sklearn 中的 train\_test\_split 方法是留出法中常用的函数，功能是从样本中随机的按比例选取 train data 和 test data。我们按照 20% 的测试集和 80% 的训练集进行划分，分离出训练集与测试集。

4.3.2 数据分析岗薪资预测神经网络模型的预训练

我们使用了 sklearn neural\_network 中的 NLPRegressor 方法进行训练，sklearn 多层感知器回归预测得到得分，并使用循环方法，遍历获取得分最高的层级。

对激活函数的选择，一般我们可以采用 sigmoid、tanh、relu 等非线性函数。sigmoid 函数取值在 图片 之间，可以将网络输出控制在上述范围以内，便于分析，但容易出现梯度消失的问题；而 tanh 函数的取值范围为 图片 之间，解决了 sigmoid 函数不是以零为中心的输出的问题，但梯度消失问题虽然得到了缓解但仍然存在。

这里最终选择使用的是 relu 函数作为激活函数，作为当今最为流行的激活函数之一，relu 函数解决了 sigmoid 和 tanh 函数计算复杂的问题，并从根本上规避了梯度消失的可能性。与 2011 年之前广泛使用的激活函数相比，可续家门在 2011 年发现它可以更好地训练更深的网络。定义 relu 函数的正部分的表达式如下：

图片

对于隐单元数目的设定，我们利用循环从 3 到 500 作为神经网络隐藏层隐单元数目，并对每个神经网络结构进行回归的预测。通过 model\_mlp.score 函数获得对当前神经网络模型的得分，并进行可视化处理，我们可以得到如下图示：

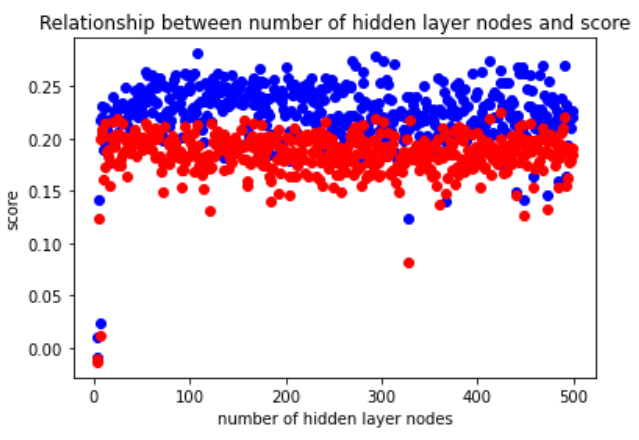


图 4.3-1 隐藏层数量与得分关系

通过第一次循环对神经网络隐藏层节点个数进行了初步筛选以后，我们发现当隐藏层的节点个数大于 50 个的时候，神经网络在训练集及测试集上的表现改进的效果并不明显，为了避免模型因为参数个数远远大于自变量的个数导致出现模型过拟合的情况，我们进一步对隐藏层节点个数为 3-50 个的情况进行了进一步的探究，结果如下：

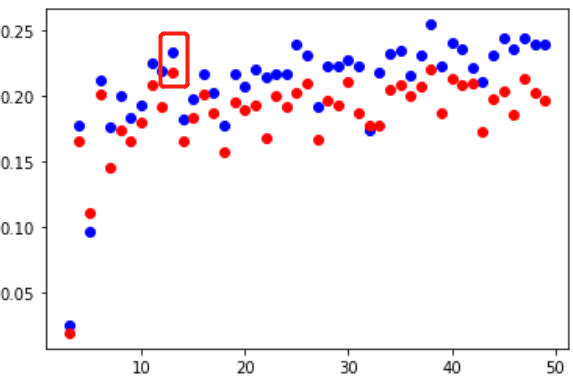


图 4.3-2 最优隐藏层单元数目

可以清晰看到，当隐藏层隐单元数目等于 13 时，模型的得分较好且模型复杂度较低，综合而言较为适合。因此，我们选择 13 个隐藏单元，至此成功建立了神经网络模型。

4.4 构建基于城市综合指标的数据分析岗薪资预测模型

在 4.3.2 节，我们利用留出法划分出训练集和测试集，对神经网络的结构进行了筛选，最终确定了使用包含一个隐藏层及 13 个隐藏节点的 BP 神经网络作为我们的最终薪资模型，并基于 4.1.3 节中的数据，使用 sklearn 包中的 neural\_network 中的 NLPRegressor 方法训练出了我们最终的薪资预测模型。输入层与隐藏层及隐藏层与输出层各节点具体的权值和偏置值见附录。

4.5 小结

本章主要介绍了基于随机森林算法的特征选择，列举了主要的回归模型并选择使用 BP 神经网络模型实现薪资预测。随后我们通过循环，比对了隐藏层数量与测试分数的关系，选择使用 13 个隐藏单元，成功建立了神经网络模型，完成了较为优秀的薪资预测。

第五章、总结与展望

5.1 主要工作总结

薪资是职场职位中的关键数据，也是吸引求职者，提高企业人才储备的重要途径。同样，对于求职者而言，薪资是决定意向的关键指标，也是决定人才流向的关键指标。

大数据环境下，互联网各式各样的求职网站数据量庞大，加之以原本的线下求职活动，诸如校招、人才市场之类的环境下，总体数据量庞大，以传统人工方式进行计算不精确，也不可取，需要我们充分发挥神经网络等计算机的特长，优化预测算法，从而进一步完善薪资预测模型的准确度与实用程度。

薪资预测模型使用的是 BP 神经网络来实现。根据循环测试，我们测试得出 13 个隐藏单元的情况下效率与准确度得分较优，从而建立了 BP 神经网络模型，验证了预测模型的有效性。

5.2 未来工作展望

本文所实现的模型还比较简单，数据特征工程处理还不够到位，因而还需要补充大量的研究和增强模型的功能，提高模型的准确率，使得它们在薪资预测中发挥更重要的作用。因此，本文未来的工作可以从以下几个方面着手:

（1）优化数据分析岗位的薪资预测模型。主要是针对 BP 神经网络模型进行改进，提高模型的预测准确性。改进的方法还有很多种，例如替换为其他机器学习算法，扩大数据集，增加样本数据特征，改进学习算法和改进误差函数等。

（2）应用更多机器学习技术到数据分析岗位薪资预测上，增强和完善系统的功能。薪资预测只是机器学习在人力资源管理系统上的初步应用，还有很多其他可以应用在不同场景下机器学习技术，因此可以实现更多有用的数据模型。

（3）优化爬虫与数据清理方式，加快数据采集与预处理时间并优化数据准确性。

（4）可以针对不同激活函数、算法和优化算法进行对比实验，探究薪资预测模型下不同算法之间组合的优劣高下，从而进一步优化薪资预测模型。

总而言之，机器学习是一门交叉范围较广的学科，广泛应用在各个领域中。将机器学习应用于数据分析岗位薪资预测中，可以为系统添加数据分析和数据预测等功能，增强薪资预测的有效性和实用性。因此，研究机器学习在薪资预测上的应用具有很好的前景，虽然现如今还只是机器学习和薪资预测方面的初步结合，但是随着时代的发展，未来还有更多的可能。

参考文献

[1] Job Salary Prediction - Kaggle https://www.kaggle.com/c/job-salary-prediction

[2] Wang Z., Sugaya S., Nguyen D.Salary Prediction using Bidirectional-GRU-CNN Model[J]. The Association for Natural Language Processing, 2019.

[3] Viroonluecha P., Kaewkiriya T. Salary Predictor System for Thailand Labour Workforce using Deep Learning[C]// The 18th International Symposium on Communications and Information Technologies. 2018.

[4] Jackman S., Reid Graham.Predicting Job Salaries from Text Descriptions[D]. University of British Columbia, 2013.

[5] 潘博, 张青川, 于重重, et al. Doc2vec在薪水预测中的应用研究[J]. 计算机应用研究, 2018, 35(001):155-157.

[6] 谷承维. 基于深度学习的垂直行业职位薪水分析与预测[D]. 北京邮电大学, 2018.

[7] 刘睿伦, 叶文豪, 高瑞卿,等. 基于大数据岗位需求的文本聚类研究[J]. 数据分析与知识发现, 2017, 001(012):P.32-40.

[8] 韦家铎. 机器学习在人力资源管理系统中的应用[D]. 华中科技大学, 2019.

致谢

寒暑两段的“英才计划”特训营转眼步入尾声，在本论文完成之际，我们首先要对我们敬爱的黎教授表示最衷心的感谢，也对一步步教导我们的施老师和刘老师表示最诚挚的感谢。感谢刘老师在授课、选题、实验与论文编写等过程中对我们的大力指导与支持。老师们认真负责，热爱数学的态度，严谨的治学态度和渊博的学识对我们影响深远，相信这些将会成为我们团队每一个人未来人生旅途上尤为珍贵的财富。

其次，我们应该感谢广州市教育局与中山大学，为我们提供了在高中阶段接触大学知识，了解大学生活的途径，为我们的计算机与数学学习提供了一个良好的学习环境。此外，我们应该感谢自诞生以来为机器学习、深度学习、神经网络做出巨大贡献的人们，是他们奠定了我们现在机器学习的基础，为我们提供了新时代信息数据处理新方法。

附录

本文所构建的基于BP神经网络的薪资预测模型的网络权值及偏置值表：

表 附录-1 网络权值表

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 0 | 0.151074 | 0.163731 | 0.262303 | -0.12429 | 0.184112 | 0.038499 | 0.084474 | 0.054444 | -0.19834 | 0.386629 | -0.07863 | 0.347168 | 0.164288 |
| 1 | -0.13541 | -0.18768 | 0.294072 | 0.202215 | 0.441207 | -0.15607 | -0.08173 | 0.006039 | -0.03539 | 0.01511 | 0.167091 | -0.29028 | -0.30409 |
| 2 | -0.30085 | -0.19299 | 0.264796 | -0.2154 | 0.006215 | -0.09318 | 0.110321 | -0.32495 | 0.374902 | 0.158265 | 0.267595 | -0.06801 | -0.34687 |
| 3 | 0.112659 | 0.212876 | 0.374617 | 0.150379 | 0.359354 | -0.14437 | -0.29955 | 0.010435 | 0.218933 | -0.17424 | -0.36987 | -0.05179 | -0.18124 |
| 4 | -0.34933 | -0.32346 | 0.231148 | 0.215556 | -0.0775 | -0.00474 | -0.09501 | -0.2484 | -0.25485 | 0.256504 | -0.18172 | -0.01421 | 0.108756 |
| 5 | 0.20204 | -0.08231 | -0.23234 | 0.346414 | -0.06068 | -0.09215 | -0.2202 | -0.25224 | -0.05792 | 0.167992 | 0.175075 | -0.26055 | -0.2602 |
| 6 | 0.374408 | -0.10311 | 0.07257 | -0.35193 | 0.010538 | -0.19611 | 0.053004 | 0.349996 | -0.0706 | -0.37541 | -0.05133 | -0.31798 | 0.084304 |
| 7 | -0.08961 | -0.0208 | 0.260769 | 0.215369 | 0.074267 | -0.31568 | -0.06898 | 0.137455 | -0.25151 | -0.34247 | -0.27478 | 0.066646 | 0.209872 |
| 8 | -0.09548 | 0.225974 | -0.03307 | 0.240949 | 0.221477 | -0.39311 | -0.21385 | 0.16144 | 0.355772 | -0.32337 | -0.3407 | -0.19165 | 0.12262 |
| 9 | -0.21719 | -0.16931 | 0.317945 | 0.033111 | -0.09533 | 0.172804 | 0.125025 | 0.092924 | -0.24212 | 0.353117 | -0.08113 | 0.089027 | 0.259077 |
| 10 | 0.14556 | -0.23698 | -0.31614 | 0.294788 | -0.2001 | 0.112028 | 0.090326 | -0.38512 | -0.31537 | -0.25029 | 0.098217 | 0.145113 | -0.0159 |
| 接上页 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10  续表 附录-1 | 11 | 12 |
| 11 | -0.26481 | -0.03679 | -0.36617 | 0.259379 | -0.27745 | -0.33553 | -0.13574 | 0.190864 | -0.26617 | 0.292435 | -0.10437 | 0.205167 | -0.01828 |
| 12 | 0.442764 | 0.173893 | -0.12545 | 0.283267 | 0.122892 | -0.07571 | -0.33919 | -0.44306 | 0.378585 | 0.251689 | -0.09063 | 0.306738 | 0.085199 |
| 13 | 0.035656 | 0.101061 | 0.210131 | -0.03855 | -0.01113 | -0.20281 | -0.29954 | -0.00708 | 0.386969 | -0.07054 | 0.075954 | -0.28558 | 0.034487 |
| 14 | 0.150488 | -0.22977 | 0.067363 | 0.243843 | 0.097431 | 0.346038 | 0.189922 | 0.131065 | -0.2829 | -0.26702 | -0.13358 | -0.03698 | -0.08749 |
| 15 | 0.320352 | 0.008378 | 0.347018 | 0.154519 | 0.191708 | 0.068847 | -0.02366 | -0.37853 | 0.183123 | 0.185057 | 0.072867 | -0.16684 | 0.238611 |
| 16 | -0.09351 | 0.114643 | -0.12927 | -0.10338 | 0.006018 | 0.045296 | 0.064991 | -0.16036 | 0.043453 | -0.15724 | -0.22358 | 0.326411 | 0.185523 |
| 17 | -0.16249 | 0.239076 | -0.19603 | -0.01272 | -0.34469 | 0.181016 | -0.08649 | 0.096228 | -0.16057 | 0.051226 | -0.35515 | -0.30447 | 0.249195 |
| 18 | 0.101048 | 0.04944 | -0.24222 | 0.098417 | -0.3247 | -0.2365 | -0.36363 | 0.009829 | 0.234521 | 0.333888 | 0.098097 | 0.316363 | -0.05458 |
| 19 | -0.28618 | -0.06737 | 0.017641 | 0.311533 | 0.441964 | 0.004139 | -0.03651 | 0.075203 | -0.18861 | -0.22392 | -0.30327 | -0.28339 | 0.305905 |
| 20 | 0.24069 | -0.00218 | 0.368655 | 0.001757 | 0.062693 | -0.14358 | -0.37884 | 0.046667 | 0.1715 | -0.23294 | -0.29824 | 0.146212 | -0.32923 |
| 21 | -0.38822 | -0.07008 | -0.1827 | -0.05937 | -0.03962 | 0.024897 | -0.19278 | 0.077292 | -0.38834 | -0.29272 | 0.145351 | -0.19007 | 0.328531 |
| 22 | -0.42204 | 0.100574 | 0.334881 | -0.39443 | -0.19436 | 0.171689 | 0.081678 | 0.155233 | 0.355774 | 0.137024 | 0.068304 | -0.30368 | 0.195534 |
| 23 | -0.04238 | -0.25487 | 0.185845 | -0.26635 | -0.12346 | 0.189926 | -0.14987 | 0.033661 | -0.09589 | -0.03322 | 0.324098 | -0.17528 | 0.11408 |

接上页

表 附录-2 偏置值表

|  |  |
| --- | --- |
|  | 0 |
| 0 | 0.012682 |
| 1 | 0.327867 |
| 2 | 0.029027 |
| 3 | 0.027434 |
| 4 | 0.10124 |
| 5 | 0.142061 |
| 6 | -0.01746 |
| 7 | 0.362756 |
| 8 | -0.01188 |
| 9 | 0.000487 |
| 10 | 0.281483 |
| 11 | -0.04516 |
| 12 | -0.04669 |