目 录

[一．任务概述 5](#_Toc19450)

[二．研究方法 5](#_Toc26406)

[2.1 LDA主题模型 5](#_Toc13148)

[2.2词频分析 5](#_Toc13243)

[2.3 线性回归 5](#_Toc2312)

[2.4 交叉验证 6](#_Toc13719)

[2.4.1 交叉验证概念 6](#_Toc23411)

[2.4.2 交叉验证目的 6](#_Toc5236)

[三．数据来源以及变量说明 6](#_Toc8743)

[四．数据分析及建模过程 6](#_Toc26012)

[4.1全国数据分析岗位探索性数据分析 6](#_Toc20865)

[4.1.1 变量说明 6](#_Toc14130)

[4.1.2 全国数据分析岗位薪资描述性统计分析 7](#_Toc25968)

[4.1.3各省份数据分析师岗位薪资分析 8](#_Toc27253)

[4.1.4 学历与数据分析师岗位薪资关系分析 10](#_Toc25363)

[4.1.5 企业规模与数据分析师岗位薪资关系分析 11](#_Toc32099)

[4.2主要城市数据分析岗位职位描述文本挖掘 12](#_Toc4202)

[4.2.1高频词统计 12](#_Toc9763)

[4.2.2 LDA主题模型 13](#_Toc23161)

[4.3上海数据分析岗位薪资预测模型构建 14](#_Toc30501)

[4.3.1变量说明 14](#_Toc25993)

[4.3.2上海数据分析岗位探索性数据分析 15](#_Toc14551)

[4.3.3上海数据分析岗位薪资预测模型构建 17](#_Toc24684)

[五．总结 21](#_Toc20059)

[参考文献： 22](#_Toc15092)

[致谢 23](#_Toc25770)

基于“数据分析师”职位信息的数据分析及建模

向丹彤

摘要：近几年，“大数据热”席卷了整个世界，市场上与大数据相关的岗位招聘、培训数不胜数。在这股数据热潮中，与数据分析相关的岗位具体情况到底如何？从企业的角度来看，数据分析岗位任职者应具备怎样的素质？对不同背景的应聘者，应如何为其定制薪资？与之相对，对于准备投身于数据科学事业的学生来说，为获得更好的职业发展前景，应重点提高自己哪方面素质？数据相关岗位就业前景究竟如何？学校又应为学生定制何种培养方案以适应市场需求？

针对这些问题，本文基于招聘网站发布的数据分析岗位的招聘信息，综合运用多种统计分析方法探究了数据分析岗位的招聘情况及薪资影响因素。从而不仅为企业招聘和薪资制定提供参考，也帮助数据相关专业学生了解岗位情况、更精准的自我定位，同时为学校的培养方案制定也起到了一定的借鉴作用。

关键词: 大数据；数据分析；统计方法；数据科学

**Data analysis and modeling based on "data analyst" position information**

Dantong Xiang

**Abstract:** In recent years ," big data fever "swept the whole world, the market and big data related to the job recruitment, training countless. In this data boom, data analysis related to the specific situation of the post after all? From the point of view of the enterprise, what qualities should the incumbent of the data analysis post have? How to customize the salary for candidates from different backgrounds? In contrast, for students who are prepared to devote themselves to the cause of data science, in order to obtain a better career development prospects, should focus on improving their own quality? What are the employment prospects for data-related jobs? What training programs should be customized for students to meet market demand? In view of these problems, this paper analyzes the recruitment information of the post based on the data published on the recruitment website, and comprehensively uses a variety of statistical analysis methods to explore the recruitment situation of the data analysis post and the factors affecting the salary. Thus not only to provide reference for enterprise recruitment and salary formulation, but also to help data-related students to understand the position situation, more accurate self-positioning, at the same time for the school training program formulation also played a certain reference role.

**Keywords**: Big Data; Data Analysis; Statistical Methods; Data Science

一．任务概述

基于招聘网站发布的数据分析岗位的招聘信息，综合运用多种统计分析方法探究数据分析岗位的招聘情况及薪资影响因素。从而不仅为企业招聘和薪资制定提供参考，也帮助相关专业学生了解岗位情况、更精准的自我定位。

二．研究方法

首先会进行相应的数据可视化。从全国范围出发，对全国数据分析岗位进行探索性数据分析。之后，将范围缩小至国内的一些主要城市，对其职位描述这一变量进行文本挖掘，主要会利用到LDA主题模型和高频词词云展示。其次，对上海数据分析岗位进行薪资预测模型构建，会考虑先利用线性回归进行建模，并对比分析其他机器学习算法在该模型的表现，使用交叉验证算法进行调参。最后，搭建了前端网页，以达到相应的展示效果。

2.1 LDA主题模型

在机器学习领域，LDA是两个常用模型的简称：Linear Discriminant Analysis 和 Latent Dirichlet Allocation。此次应用的Latent Dirichlet Allocation. LDA 在主题模型中占有非常重要的地位，常用来文本分类。LDA由Blei, David M.、Ng, Andrew Y.、Jordan提出，用来推测文档的主题分布。LDA是文本语义分析中应用广泛的一个模型，比如找相似文档等应用场景，如果知道每篇文档的主题分布，那么根据隐含的主题分布而不是简单的词统计数据来计算文档的相似度等。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题分布后，便可以根据主题分布进行主题聚类或文本分类。

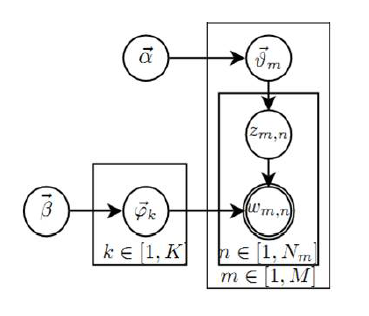


图0.1 LDA原理图

2.2词频分析

词频分析是对文献正文中重要词汇出现的次数进行统计与分析，是文本挖掘的重要手段。它是文献计量学中传统的和具有代表性的一种内容分析方法，基本原理是通过词出现频次多少的变化，来确定热点及其变化趋势。

2.3 线性回归

线性回归是一种通过属性的线性组合来进行预测的线性模型，其目的是找到一条直线或者一个平面或者更高维的超平面，使得预测值与真实值之间的误差最小化。线性回归得出的模型不一定是一条直线，在只有一个变量的时候，模型是平面中的一条直线；有两个变量的时候，模型是空间中的一个平面；有更多变量时，模型将是更高维的。线性回归模型有很好的可解释性，可以从权重W直接看出每个特征对结果的影响程度。线性回归适用于X和y之间存在线性关系的数据集，可以使用计算机辅助画出散点图来观察是否存在线性关系。看起来这些点分布在一条直线附近，尝试使用一条直线来拟合数据，使所有点到直线的距离之和最小。实际上，线性回归中通常使用残差平方和，即点到直线的平行于y轴的距离而不用垂线距离，残差平方和除以样本量n就是均方误差。均方误差作为线性回归模型的代价函数。

2.4交叉验证

2.4.1 交叉验证概念

交叉验证主要用于建模应用中，例如 PCR 、PLS 回归建模中。在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。

2.4.2 交叉验证目的

用交叉验证的目的是为了得到可靠稳定的模型。用交叉校验每个主成分下的PRESS值，选择PRESS值小的主成分数。或PRESS值不再变小时的主成分数。常用的精度测试方法主要是 交叉验证，例如10折交叉验证，将数据集分成十份，轮流将其中9份做训练1份做测试，10次的结果的均值作为对算法精度的估计，一般还需要进行多次10折交叉验证求均值

三．数据来源以及变量说明

本文所用数据均爬取自Boss直聘网站（https://www.zhipin.com/）。其中包括全国范围数据近60,000条，为全国数据分析岗位信息的探索性数据分析及主要城市职位描述文本挖掘部分用；上海范围数据为构建上海数据分析岗位的薪资预测模型用。

全国数据分析岗位数据包括以下变量：公司名称、公司规模、省、城市、职位名称、职位描述、工作经验、学历。

上海数据分析岗位的数据主要包括以下变量：公司规模、融资情况、区、工作经验、学历、技能。

四．数据分析及建模过程

4.1全国数据分析岗位探索性数据分析

4.1.1 变量说明

本部分包括以下变量：公司名称、公司规模、省、城市、职位名称、职位描述、工作经验、学历。具体描述如表1.1所示。

表1.1 全国数据分析岗位变量描述

|  |  |  |  |
| --- | --- | --- | --- |
| 变量类型 | 变量名 | | 详细说明 |
| 因变量 | 平均薪资 | | 连续变量，单位：k |
| 自变量 | 企业因素 | 公司名称 | 字符型变量 |
| 公司规模 | 定性变量，共6个水平  （0-20人、20-99人、100-499人、500-999人、1000-9999人、10000人以上） |
| 省 | 定性变量，共28个水平 |
| 城市 | 定性变量，共335个水平 |
| 职位名称 | 定性变量  （例：数据分析师、风险数据分析师） |
| 职位描述 | 字符型变量 |
| 个人因素 | 工作经验 | 定性变量，共7个水平  （经验不限、应届生、1年以内、1-3年、3-5年、5-10年、10年以上） |
| 学历 | 定性变量，共9个水平  （学历不限，初中及以下，中专及以下，中专/中技，高中，大专，本科，硕士，博士） |

4.1.2 全国数据分析岗位薪资描述性统计分析

绘制全国数据分析岗位薪资分布直方图，并计算其均值、标准差、最小值、最大值、四分之一分位点、中位数、四分之三分位点。结果分别示与图1.1，表1.2中。

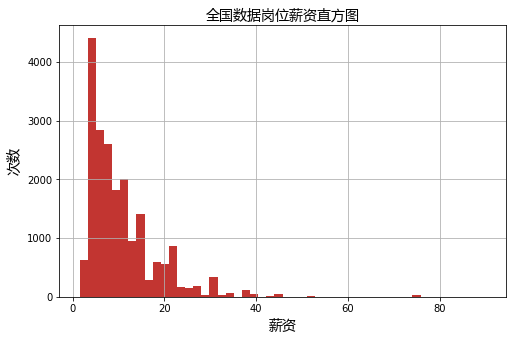
 图1.1全国数据分析岗位薪资直方图

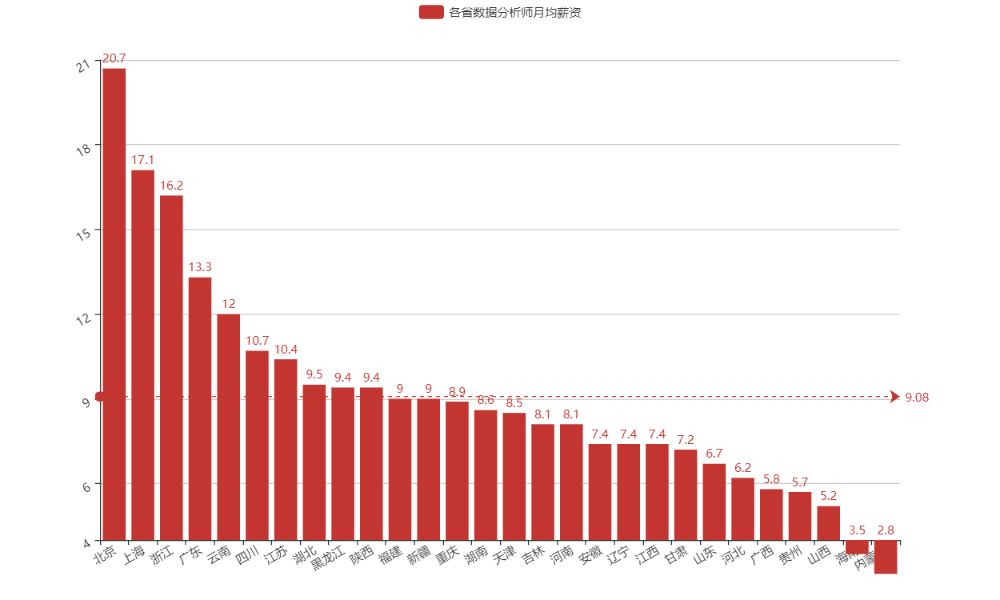
表1.2 全国数据分析岗位薪资描述性统计分析

|  |  |
| --- | --- |
| 统计量 | 数值（单位：千） |
| 平均数 | 10.7 |
| 标准差 | 7.7 |
| 最小值 | 1.5 |
| 25%分位数 | 5.5 |
| 中位数 | 8.0 |
| 75%分位数 | 13.5 |
| 最大值 | 90.0 |

由表1.2可知，全国数据分析岗位薪资平均水平在10700左右，最大值为90000，远远超过1.5倍的75%分位数与25%分位数之差，但考虑到中国城市区域发展不均衡的情况，在此部分不考虑将其列为异常值进行处理。

4.1.3各省份数据分析师岗位薪资分析

将数据分析师岗位薪资按省份分组求平均值，其结果见图1.2所示。

****图1.2 各省份数据分析岗位月均薪资

由图1.2可知，全国各省份数据分析岗位月均薪资均值仅为9.08k，相对来说比较低。综合全国来看，北京、上海、浙江、广东等发达地区的月均薪资远远高于全国平均水平，其中北京高达20.7k。但广西、贵州、山西、海南、内蒙古等较为偏远地区的月均薪资普遍偏低，内蒙古仅有2.8k，比北京低18k左右。造成如此显著差异的原因主要是我国经济发展不均衡，各地平均收入和消费水平本身就差异较大。另一方面，这一现象也反映了我国高素质人才的流动方向，“数据分析师”这一岗位主要存在于大中型企业和科技型企业中，因此发达地区对这些人才的需求较大，也吸引更多高素质人才涌入北京、上海等城市。图1.3更加直观的展示了结果。



图1.3 全国数据分析师薪资分布图

由图1.3可更加直观的获得全国数据分析师薪资分布情况，除北京、上海等少数地区外，大部分省份的数据分析师薪资处于较低水平，这意味着我国数据分析师岗位市场仍具有大的挖掘潜力。另外，本文在爬取数据过程中并未获取到西藏、青海、宁夏的有效信息，其原因为这三个省份在全国处于经济发展落后水平，其并无数据分析师相关岗位设置。

4.1.4 学历与数据分析师岗位薪资关系分析

运用箱线图分析学历与薪资间的关系，结果见图1.4所示，其中箱体宽度表示岗位数。

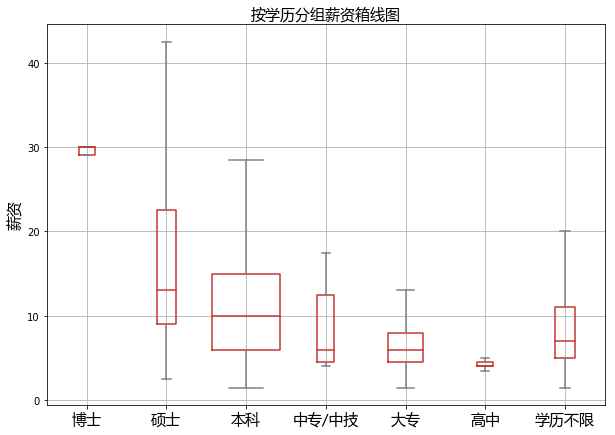
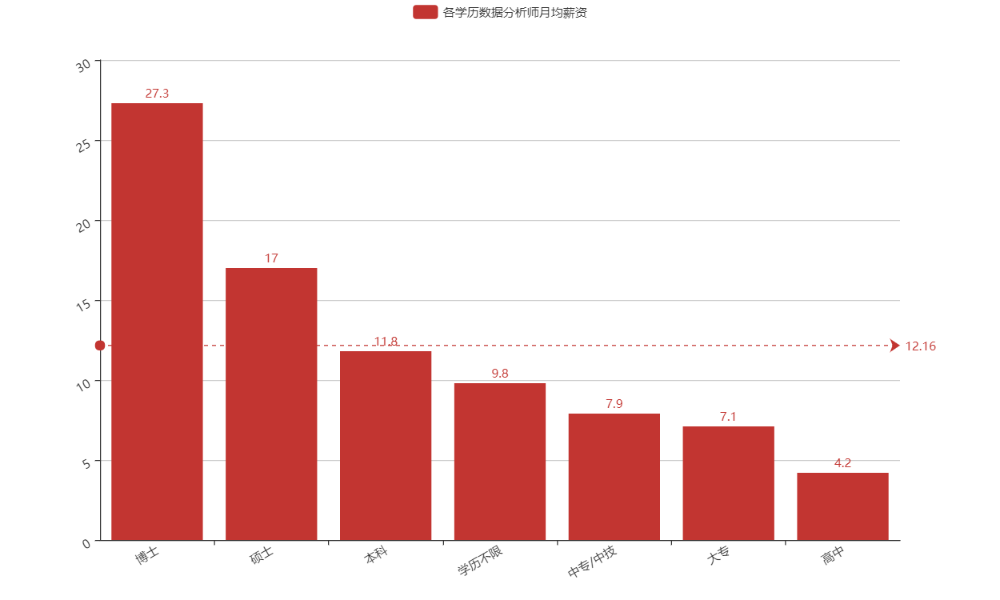


图1.4 按学历分组数据分析师岗位薪资箱线图

从图1.4中箱体水平宽度可知，数据分析岗位大多要求中等学历，本科最多，要求硕博士或高中及以下的岗位较少。学历要求为博士的岗位数目最少，但薪资显著高于其他岗位。图1.5为各学历水平平均月薪情况。

图1.5 各学历数据分析师月均薪资分布情况

由图1.5，薪资随学历提高而不断上升，博士薪资远高于其他学历。博士、硕士、本科间的月均薪资差距明显。而本科以下学历所得月均薪资差距不明显。

4.1.5 企业规模与数据分析师岗位薪资关系分析

将数据分析师岗位薪资按照企业规模分组求均值，结果见图1.6。

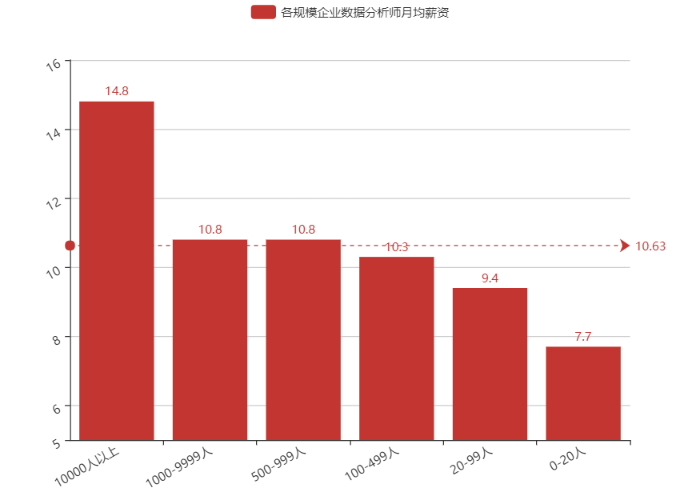


图1.6 各规模企业分析师月均薪资分布情况

由图1.6可知，企业规模为10000人以上的大型企业数据分析师月均薪资为14.8k，远高于其他中小型企业。而对于中小型企业，其企业规模对数据分析师薪资影响不明显。

4.2主要城市数据分析岗位职位描述文本挖掘

本部分基于对全国范围数据分析岗位探索性数据分析结果，综合各方因素，选出五个主要城市进行职位描述信息的文本挖掘。主要城市包括：北京、上海、广州、深圳、杭州。

4.2.1高频词统计

对上述城市数据分析岗位的职位描述文本进行分词，并统计高频词。词频位于前30的关键词及其词频见表2.1。

表2.1 主要城市职位描述词频TOP30

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 关键词 | 词频 | 序号 | 关键词 | 词频 | 序号 | 关键词 | 词频 |
| 1 | 经验 | 2580 | 11 | 工具 | 737 | 21 | 支持 | 578 |
| 2 | 沟通 | 1064 | 12 | 用户 | 736 | 22 | 建模 | 573 |
| 3 | 报告 | 1017 | 13 | SQL | 703 | 23 | 监控 | 563 |
| 4 | 团队 | 1010 | 14 | 挖掘 | 695 | 24 | 策略 | 562 |
| 5 | 以上学历 | 950 | 15 | 本科 | 677 | 25 | 建议 | 544 |
| 6 | 行业 | 880 | 16 | 模型 | 664 | 26 | 研究 | 544 |
| 7 | 问题 | 879 | 17 | 统计学 | 640 | 27 | 平台 | 519 |
| 8 | 统计 | 863 | 18 | 互联网 | 617 | 28 | 体系 | 502 |
| 9 | 数学 | 785 | 19 | 开发 | 588 | 29 | 逻辑 | 489 |
| 10 | 数据挖掘 | 748 | 20 | 能够 | 586 | 30 | 算法 | 488 |

“经验”排在第一位，这说明数据分析岗位对经验要求很高，具有一定经验的数据分析师更加受招聘企业欢迎。而其他排名较高的词语，如“沟通”、“团队”等，则是各个行业、各个岗位都必须具备的素质。另外，一些能体现数据分析岗位工作内容和特点的词语，如“统计”、“数学”、“数据挖掘”、“SQL”、“建模”等，也都位于前列。基于对高频词词频统计，本文绘制了词频位于前200的高频词词云图，如图2.1所示。

图2.1 主要城市职位描述词频TOP200词云图

4.2.2 LDA主题模型

4.2.2.1 LDA主题模型原理简介

LDA由Blei等人于提出，用来推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题分布后，便可以根据主题分布进行主题聚类或文本分类。

其基本思想为：假定语料库中有*m*篇文档，共涉及*k*个主题，每个主题下的词服从参数为的Dirichlet先验分布中采样得到的多项分布，每篇文档长度为。从一参数为的Dirichlet先验分布中中采样得到一个多项分布作为该文档中每个主题的概率分布；对于某篇文档中的第*n*个词，首先从该文档中出现每个主题的多项分布中采样一个主题，然后再在这个 主题对应的词的多项分布中采样一个词。不断重复这个随机生成过程，直到*m*篇文档全部完成上述过程。

4.2.2.2 LDA主题模型构建结果

使用LDA主题模型，对职位描述文本进行主题划分，得到五个主题及其对应关键词，结果见表2.2。

表2.2 职位描述LDA主题模型结果

|  |  |  |
| --- | --- | --- |
| 序号 | 主题 | 关键词 |
| 1 | 游戏运营 | 游戏、监控、报表、开发 |
| 2 | 风险战略 | 金融、策略、商业、建议 |
| 3 | 算法开发 | 算法、模型、研究、挖掘 |
| 4 | 用户导向 | 用户、团队、建模、报告 |
| 5 | 互联网 | 互联网、开发、统计、行业 |

由表2.2中结果可得，五个主题模型反映了数据分析岗位的五个主要的细分业务发展方向：游戏运营、风险战略、算法开放、用户导向、互联网。这一结果与目前数据分析岗位市场状况相符合。

4.3上海数据分析岗位薪资预测模型构建

4.3.1变量说明

本部分包括以下变量：公司规模、融资情况、区、工作经验、学历、技能。具体变量描述见表3.1。

表3.1 上海数据分析岗位变量描述

|  |  |  |  |
| --- | --- | --- | --- |
| 变量类型 | 变量名 | | 详细说明 |
| 因变量 | 平均薪资 | | 连续变量，单位：k |
| 自变量 | 企业因素 | 公司规模 | 定性变量，共6个水平  （0-20人、20-99人、100-499人、500-999人、1000-9999人、10000人以上） |
| 融资情况 | 定性变量，共14个水平  （A轮、B轮、C轮、D轮及以上、天使轮、未融资、不需要融资、已上市、0-20人、20-99人、100-499人、500-999人、1000-9999人、10000人以上） |
| 地区 | 定性变量，共15个水平 |
| 个人因素 | 学历 | 定性变量，共9个水平  （学历不限，初中及以下，中专及以下，中专/中技，高中，大专，本科，硕士，博士） |
| 工作经验 | 定性变量，共7个水平  （经验不限、应届生、1年以内、1-3年、3-5年、5-10年、10年以上） |
| 技能 | 定性变量，共10个水平  （Python、SQL、R、Java、Hadoop、C、Spark、Excel、Hive、Sas） |

4.3.2上海数据分析岗位探索性数据分析

4.3.2.1上海数据分析岗位薪资描述性统计分析

绘制上海数据分析岗位薪资分布直方图，并计算其均值、标准差、最小值、最大值、四分之一分位点、中位数、四分之三分位点。结果分别示与图3.1，表3.2中。

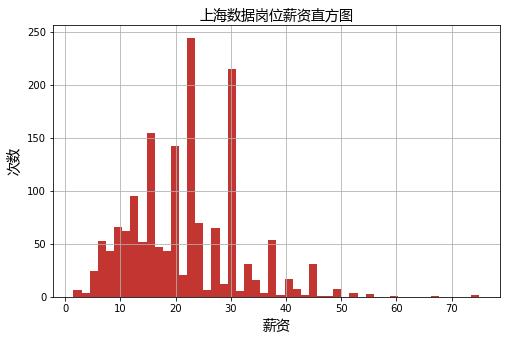
图3.1 上海数据分析岗位薪资直方图

表3.2上海数据分析岗位薪资描述性统计分析

|  |  |
| --- | --- |
| 统计量 | 数值（单位：千） |
| 平均数 | 21.5 |
| 标准差 | 9.9 |
| 最小值 | 1.5 |
| 25%分位数 | 14.5 |
| 中位数 | 22.0 |
| 75%分位数 | 29.5 |
| 最大值 | 75.0 |

由表3.2结果可知，上海数据分析岗位平均月薪为21.5k，最大值75.0k，最小值1.5k，薪资差距依旧十分明显。同时，标准差高于全国水平，说明上海薪资波动情况相较全国来说更为剧烈。

4.3.2.2软技能分析

软技能是指任职者的无法量化衡量的综合能力。匹配上海地区数据分析岗位职位描述文本中出现的软技能，并统计其频数。若一个岗位对应职位描述中某一关键词多次出现，只按一次计算。软技能词频统计结果示于表3.3中。

表3.3上海职位描述软技能词频统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 软技能 | 词频 | 序号 | 软技能 | 词频 |
| 1 | 数据分析 | 861 | 6 | 统计 | 505 |
| 2 | 数学 | 697 | 7 | 建模 | 488 |
| 3 | 算法 | 686 | 8 | 报告 | 386 |
| 4 | 数据挖掘 | 661 | 9 | 商业 | 384 |
| 5 | 机器学习 | 625 | 10 | 数据库 | 285 |

“数据分析”是一较为概括和综合的技能概念，也与数据分析岗位密切相关，位列软技能第一位。另外，数据分析岗位也要求应聘者具有较好的数学功底及基本编程技能，其中“数学”、“算法”、“数据挖掘”、“机器学习”也是排在前五位。

4.3.2.3硬技能分析

硬技能与软技能相对，指从事某一工作的硬性要求，与软技能互补，都是应聘者必须具备的素质。匹配上海地区数据分析岗位职位描述文本中出现的硬技能，并统计其频数。计数方式同软技能，硬技能词频统计结果示于表3.4中。

表3.4上海职位描述硬技能词频统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 硬技能 | 词频 | 序号 | 硬技能 | 词频 |
| 1 | Python | 834 | 6 | C | 276 |
| 2 | SQL | 592 | 7 | Spark | 263 |
| 3 | R | 480 | 8 | Excel | 258 |
| 4 | Java | 335 | 9 | Hive | 239 |
| 5 | Hadoop | 303 | 10 | Sas | 218 |

词频位于前三的Python、SQL、R都是数据分析岗位最常使用的工具。而基础编程语言Java、C排名居中，分别位于第4名、第6名。而Sas这一传统统计工具，由于其应用范围的局限性，排名最末，在业内并无广泛应用。

计算上海数据分析岗位硬技能与薪资相关系数，探究其关系，结果列于表3.5中。

表3.5上海数据分析岗位硬技能与薪资相关系数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 硬技能 | 相关系数 | 序号 | 硬技能 | 相关系数 |
| 1 | Java | 0.261871 | 4 | Python | 0.189326 |
| 2 | Spark | 0.221214 | 5 | Hadoop | 0.182821 |
| 3 | C | 0.213868 | 6 | Hive | 0.172596 |

Java与薪资相关系数最高，但结合表3.4结果，Java在硬技能词频排序中仅位于第四，这意味着Java与薪资关系密切，掌握Java很大程度上意味着高薪，但这并不是数据分析岗位的普遍要求。Python与其情况相反，Python这一技能在词频排序中靠前但与薪资相关系数并不是很高，这说明Python几乎是数据分析岗位的必备技能。

4.3.3上海数据分析岗位薪资预测模型构建

本文在此部分针对表3.1中变量建立薪资预测模型。模型构建简要思路如下：首先，采用原始特征训练基础模型及机器学习模型；之后，加入衍生特征进行机器学习模型训练；接着，选取表现较好的模型进行调参；最后，将调参后的模型作为堆叠模型的基学习器。

4.3.3.1采用原始特征训练模型

首先考虑使用线性回归作为基础模型，本文使用线性回归及其衍生模型，包括对数线性回归、Lasso、主成分回归进行模型构建，各基础模型测试误差列于表3.6中。

表3.6 基础模型算法总结

|  |  |  |
| --- | --- | --- |
| 序号 | 模型 | 测试误差 |
| 1 | 线性回归 | 7.53694105153 |
| 2 | 对数线性回归 | 7.68995072203 |
| 3 | Lasso | 7.52938936207 |
| 4 | 主成分回归 | 7.48123926645 |

其中，主成分回归测试误差最小，表现最好。将主成分回归模型与其他回归算法进行比较，包括决策树回归、k邻近回归。引入集成学习，建立随机森林模型，Adaboost模型，GBDT模型、Bagging模型以及Xgboost模型，通过训练误差和测试误差比较其表现。结果如表3.7所示。

表3.7 机器学习算法总结

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 模型 | 训练误差 | 测试误差 |
| 1 | PCA Reg | 7.72050419197 | 7.48123926645 |
| 2 | KNN | 6.75157708269 | 7.74633925232 |
| 3 | Decision Tree | 1.75392511359 | 9.19670236737 |
| 4 | Bagging | 3.81819552578 | 7.54527836246 |
| 5 | Adaboost | 6.75157708269 | 8.95363465919 |
| 6 | GBDT | 6.86403850991 | 7.17821600008 |
| 7 | Random Forest | 3.74829864056 | 7.43739341558 |
| 8 | Xgboost | 3.85148621231 | 7.35382054896 |

综合表3.6和表3.7结果，本文选择Lasso、随机森林、Xgboost、GBDT进行调参并构建堆叠模型。

4.3.3.2特征工程

考虑对特征进行进一步处理。具体处理过程如下：

（a）去除特征重要性为0的变量，测试误差降为7.35382054896。

（b）根据类别变量分组后计算薪资平均值进行类别变量编码，同时采用后退法筛选特征。得到类别变量包括公司规模、区、融资情况、学历要求。此时，测试误差降为7.25712440821。

（c）采用DBSCAN法进行聚类，增加聚类标签。

DBSCAN算法是在1996年由Ester等人提出的以密度为本的聚类分析算法。其基本思想为：首先任意选择一个没有类别的核心对象作为种子，找到所有这个核心对象能够密度可达的样本集合，即为一个聚类簇；继续选择另一个没有类别的核心对象去寻找密度可达的样本集合，得到另一个聚类簇。重复迭代直到所有核心对象都有类别。

首先，增加个人因素聚类标签，个人因素包括工作经验、学历要求，测试误差降为4.18407616321。之后，增加公司因素聚类标签，公司因素包括公司规模、地区、融资情况，此时测试误差降为3.61681988681。最后，增加技能聚类标签，技能包括Java、Python、SQL等，此时测试误差为3.70012354984，效果反而变差，故不考虑技能聚类标签。

上述特征工程汇总至表3.8。最终选择去除特征重要性为0的特征，增加五个类别变量编码、个人因素聚类标签、公司因素聚类标签。

表3.8 特征工程总结

|  |  |
| --- | --- |
| 方法 | 测试误差 |
| 去除特征重要性为0的特征 | 7.35382054896 |
| 增加类别变量编码同时用后退法筛选特征 | 7.25712440821 |
| 个人因素聚类标签 | 4.18407616321 |
| 个人因素聚类标签+公司因素聚类标签 | 3.61681988681 |
| 技能聚类标签+个人因素聚类标签+公司因素聚类标签 | 3.70012354984 |

4.3.3.3调参

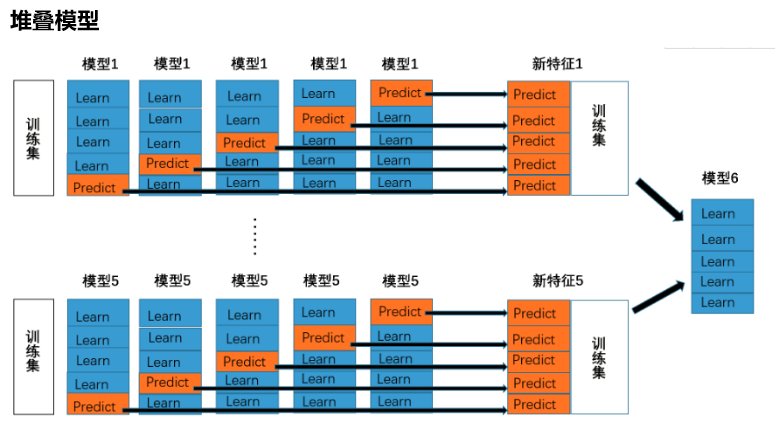
根据特征工程结果，并选择Lasso、随机森林、Xgboost、GBDT进行调参。调参结果列于表3.9中。

表3.9 调参结果

|  |  |  |
| --- | --- | --- |
| 模型 | 参数 | 最优 |
| Lasso | alpha | 0.1 |
| Random Forest | n\_estimators | 150 |
| max\_depth | 25 |
| min\_samples\_split | 2 |
| Xgboost | n\_estimators | 900 |
| max\_depth | 3 |
| min\_child\_weight | 5 |
| reg\_alpha | 1e-05 |
| GBDT | min\_samples\_split | 500 |
| max\_depth | 20 |
| n\_estimators | 624 |

4.3.3.4堆叠模型

堆叠（也称为元组合）是用于组合来自多个预测模型的信息以生成新模型的模型组合技术。即首先训练多个不同模型，之后以其输出为输入来训练一个模型，以得到一个最终的输出。其原理简要示意图如图3.2所示。

图3.2 堆叠模型原理示意图

本文将原始训练集划分为5折大小相同且互不相交的子集，以调参后的Lasso、Xgboost、GBDT、随机森林模型为弱学习器。对于每一个弱学习器，重复k次学习，每次学习过程以K-1折子集作为训练集，预测剩余的第k个子集。重复训练四个模型，基于预测所得数据结合原始标签以得到新的训练集，通过新的训练集学习元回归器，从而得到最终模型。堆叠模型总结如表3.10所示。

表3.10 堆叠模型总结

|  |  |  |
| --- | --- | --- |
| 学习器 | | 参数 |
| 基学习器 | Lasso | alpha=0.1 |
| Xgboost | n\_estimators=900, max\_depth=3, min\_child\_weight=5, reg\_alpha=1e-5 |
| GBDT | min\_samples\_split=500, max\_depth=20, n\_estimators= 624 |
| Random Forest | n\_estimators=150, max\_depth=25, min\_samples\_split=2 |
| 元学习器 | Linear Regression | 无 |

堆叠模型最终测试误差降至3.5625312428187335。

4.3.3.5薪资预测

基于上述所构建模型，建立薪资预测器，方便应聘者进行职业测评。部分结果示于表3.11中。另外，本文在此部分搭建了简易网页为更好的展示成果。

表3.11 薪资预测示例

|  |  |  |
| --- | --- | --- |
| 序号 | 自我测评因素 | 预测薪资（单位：千） |
| 1 | 公司规模：10000人以上，学历要求：硕士  工作经验：1年以内，技能：Python、SQL、R | 18.58 |
| 2 | 公司规模：10000人以上，学历要求：本科  工作经验：1-3年，技能：Java、Python、Hadoop | 20.85 |
| 3 | 公司规模：100-499人，学历要求：硕士  工作经验：1年以内，技能：Java、Python、Hadoop、R | 21.96 |

五．总结

本文仍存在很多缺点和不足，并对此提出后续研究方向：

1.本文所选取变量仅限于从BOSS直聘网上能直接获取到的信息，并未挖掘更多能够对薪资造成影响的因素及隐含信息进行分析。

2.由于BOSS直聘网提供的薪资为一个区间，本文对此的处理方式为取其平均值，但实际上应聘者真实获得的薪资通常低于均值，故本文在后续的研究中考虑对薪资的量化采用更加精确完善的处理方式。

3.本文在构建聚类标签时仅采用了基于密度的DBSCAN法，后续考虑尝试其他聚类方法，如K-means聚类法、层次聚类法等。

4.本文在构建薪资预测模型及网页搭建时，仅考虑了自我测评因素，在后续研究中，可考虑加入职位描述相关信息，根据职位描述信息进行薪资预测。

参考文献：

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2012, 3:993-1022.

[2] Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery & Data Mining. 1996.

[3] Sander J , Ester M , Kriegel H P , et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications[J]. Data Mining and Knowledge Discovery, 1998, 2(2):169-194.

[4] <http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/>

致谢

本论文是在闫雷鸣老师的亲切关怀和悉心指导下完成的,他严肃的科学态度,严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。闫雷鸣老师不仅在学业上给我以精心指导，同时还在思想、生活上给我们以无微不至的关怀，在此谨向闫雷鸣老师致以诚挚的谢意和崇高的敬意。在课设即将完成之际，我的心情无法平静，从开始进入课题到课设论文的顺利完成，有多少可敬的老师、同学、朋友给了我无言的帮助，在这里请接受我诚挚的谢意!最后，再次对关心帮助我的老师和同学表示衷心地感谢!