人才吸引力评价模型研究

摘要

本文基于地区人才吸引力要素设计调查问券，并添加工作年限与工作类型。然后通过包裹式特征过滤与本文设置的误差过滤标砖进行特征过滤，通过对人才的整体，分类型，分工作年限进行Logistics回归分析，从而实现了对人才的分类型以及动态分析。

基于第一问，为了建立济南人才吸引力模型，设计了济南四维度吸引力要素分析框架。然后搜集数据，并利用包裹式特征过滤选择与建立的交叉验证误差标准阈值选择特征。最后对数据进行回归分析得到模型：

根据济南市人才吸引模型权重与济南市公共财政收入增长得出，济南放管服政策提高了济南市19.115%的人才吸引力。

基于第二问，为了对比济南与同类型城市天津对不同类型人才吸引力的差异，并给出相应的施政方案。本文根据科研教育，研发开发，金融管理三类人才进行了不同的Logistics回归分析。然后对比天津市人才吸引模型。展示了两市之间对不同类型人才吸引力的差距，并根据现有状况提出了进行收入补贴与住房补贴消除抵消住房条件与收入水平负面影响与扩大家庭保障维度与生活质量维度优势的相关政策。

基于第三问，为了能够动态分析人才的不同需求，根据人才的工作年限分别进行了Logistics回归分析。得到了人才在不同发展时期的动态需求，并以此对人才的前期发展提出降低迁移成本，提供住房补助，保障家庭生活质量的具体措施，对处于中期发展的人才提出向下延伸贡献奖励金分档，增加假期，旅游补助等具体措施提高事业维度与生活质量维度吸引力。对于处于发展后期的人才，则提出一系列能够使其更好享受济南城市生活的便利与舒适的措施。

**关键词**：人才吸引力要素; 包裹时特征选择 ;Logistics回归分析;交叉验证

1. 问题重述
   1. 问题背景

21世纪最稀缺的是人才，一个城市要保持其竞争力和创新力，人才吸引是政府工作的重中之重。2018年济南为吸引高新企业和优秀人才，促进新能源的发展，把加大营商环境改力度作为年度工作重点。

“迎合人才的理想，满足人才需求“是吸引人才的关键。对于大多数人才来说，最关心的还是”发展前景“：行业的发展前景和城市的发展前景，不止要当前好，未来发展也要一片光明，因为“迁移成本”也是人才的考虑因素之一；再者是收入，（薪水或盈利），这方面有绝对（同行业）和相对（同地域，平价购买力）两种考量因素；再次是环境方面的因素：交通、教育、医疗、购物、治安等等。目前研究，这方面定量研究少，定性讨论多；定量研究中也是单因素多，综合因素少；动态考量少，静态考量多，能考虑到“不可比”条件的更少。主要原因还是缺乏合适的数学模型，结论缺乏说服力和可验证性。

1.2 问题提出

**问题1** 收集数据来建立数学模型，以此量化评价济南市的人才吸引力水平，就“改善营商环境 济南打造“放管服”改革样本”对济南的人才吸引力水平做出量化评价。

**问题2** 针对人才不同类别，进行深入分析，比较济南市与其他同类城市（如天津、青岛、苏州等）在人才吸引力水平的优势与不足，给出有效的提升人才吸引力的可行方案。

**问题3** 针对济南市高新区的技术发展特点和人才政策，并考虑人才的多方面需求，量化评价济南市高新区人才吸引力水平。

1. 问题分析

基于问题1，为了得到济南市人才吸引力模型。首先根据题目要求与以往的地区人才吸引要素研究，寻找合适的地区人才吸引要素，最终建立了四维度的地区人才吸引力要素框架，其中：基础环境维度三个： 基础设施建设, 生态环境, 治安环境；事业环境维度四个： 产业发展, 企业条件, 职业发展, 收入水平；家庭保障维度四个： 住房条件, 户籍政策, 配偶就业,子女教育；生活维度三个：生活成本, 交通状态, 休闲娱乐。并有性别，地区，月收入，工作性质，工作年限四个额外问题为之后的问题做准备。然后通过QQ,微信等多种途径,呼吁与发放调查问卷链接最终得到202份调查问卷。其中来自山东的问卷79份，其余问卷则来自全国各地。然后对获取的数据进行数据处理，利用包裹式特征选择（逻辑回归位基学习器）对特征进行选择，设立了误差阈值，最终确定所有特征均在误差阈值内。然后利用Logistics逻辑回归对数据进行回归分析，得出公式

济南市放管服改革作用与事业维度的产业发展，若依据2017年第一季度地方公共财政收入同比增加31.7%计量，那么放管服改革在2016第一季度到2017第二季度使济南市的人才吸引力提高了19.115%。

基于问题2，我们将济南与同类型地区天津进行深入对比，从人才引进类型的角度来看不同类型的人才受不同地区人才吸引要素的影响方式与影响力并不相同。首先通过对不同类型的人才分开进行建模得到不同类型人才的地区吸引力模型。济南，天津人才引进共分为三个大类：一，科研教育类，包括科学家，大学教授，学术带头人等；二，技术研发类，包括机械与电子产业的工程师，设计师，技术顾问等，三，管理销售类，高级管理人才与销售人才，此类人才在数据分布上具有较强的一致性。借鉴与以往对天津地区人才吸引力的研究，对比天津两市对不同类型人才吸引力状况，提出进一步促进产业发展，为高层次人才提供奖金与住房补助等有效措施。

基于问题3，根据济南市高新区的人才需求,从数据中筛选出金融从业者,技术研发人才,技工人才,管理人员,教育科研五个工作方向的数据,然后根据其工作年限将其划分为短期(1年工作经验及以下),中长期(2-5年工作经验),长期(6年以上工作经验).然后利用Logistics逻辑回归进行回归分析.得到人才在不同时期不同的发展需求。人才在工作初期更在意休闲娱乐，户籍政策与产业发展，子女教育，职业发展五个因素，因此对于刚工作的人才应当降低其迁移成本，提高其生活质量与家庭保障。而对于子女教育，生活成本，休闲娱乐，交通状态与生态环境五个维度，因此应当更好的解决其子女教育问题，并为其生活质量提供保障与支持。对于工作时间已经较长的人才而言，其更注重基础设施建设，收入水平，交通状态以及产业发展。因此对于已经有充足工作经验的人才，应当从公共状况与产业发展角度入手，更好的使其感到便利。

1. 模型假设
2. 假设调差问卷的数据分布与济南人才需求分布差异对最终结果无显著影响。
3. 假设调差问卷的数据不存在重复调查问题。
4. 假设调查问卷中被调查者的回答如实可靠。
5. 符号说明

|  |  |  |  |
| --- | --- | --- | --- |
| **符号** | **符号说明** | **符号** | **符号说明** |
| **Y1** | 济南是否具有吸引力 | **Bas** | 基础设施建设 |
| **Y2** | 济南是否对技术研发类人才具有吸引力 | **Env** | 生态环境 |
| **Y3** | 济南是否对  管理销售类人才具有吸引力 | **Sec** | 治安环境 |
| **Y4** | 济南是否对  科研医疗类人才具有吸引力 | **Ind** | 产业发展 |
| **Y5** | 济南是否对  其他人才具有吸引力 | **Comenv** | 企业条件 |
| **Y6** | 济南是否对  工作年限为 1年及以下人才具有吸引力 | **Occ** | 职业发展 |
| **Y7** | 济南是否对  工作年限为 2-3年与4-5年人才具有吸引力 | **Inc** | 收入水平 |
| **Y8** | 济南是否对  工作年限为 6-9年与10年及以上人才具有吸引力 | **Hou** | 住房条件 |
| **Sigmoid** |  | **Cen** | 户籍政策 |
| **Tran** | 交通状态 | **Spou** | 配偶就业 |
| **Ente** | 休闲娱乐 | **Edu** | 子女教育 |
| **Cost** | 生活成本 |  |  |

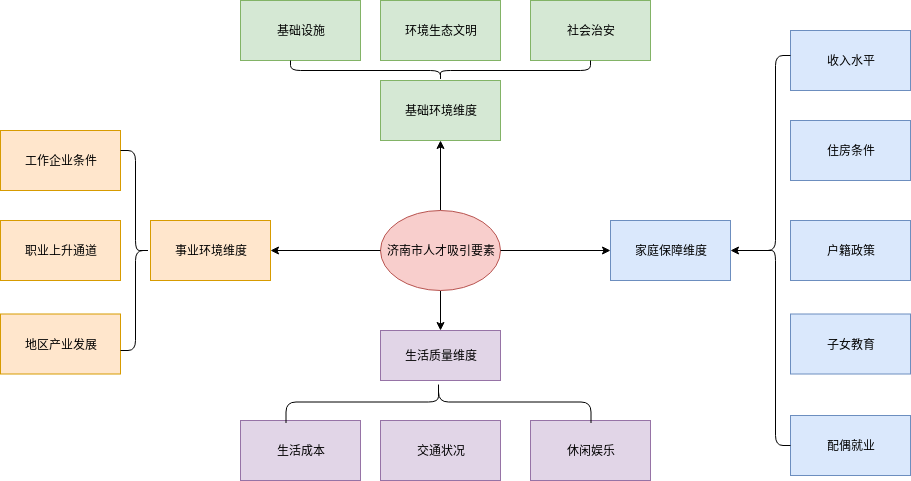
1. 模型的建立和求解

**5.1问题一模型的建立与求解**

**5.1.1建模前的准备**

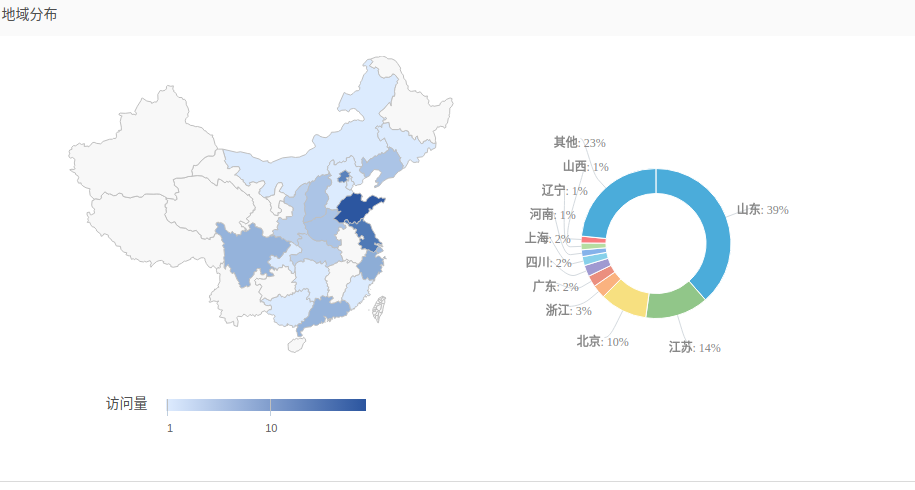
（1）建立济南地区人才吸引力要素框架

根据传统地区人才吸引力的有关研究建立的四维度整体分析框架进行设计，并参考了张炜在天津市人才吸引力所使用的四维度分析框架，对之在做出少量修改的同时并增加了人才类别与工作年限等其他重要因素的统计，以便于获取更为有价值的信息。



1. 发放调查问卷

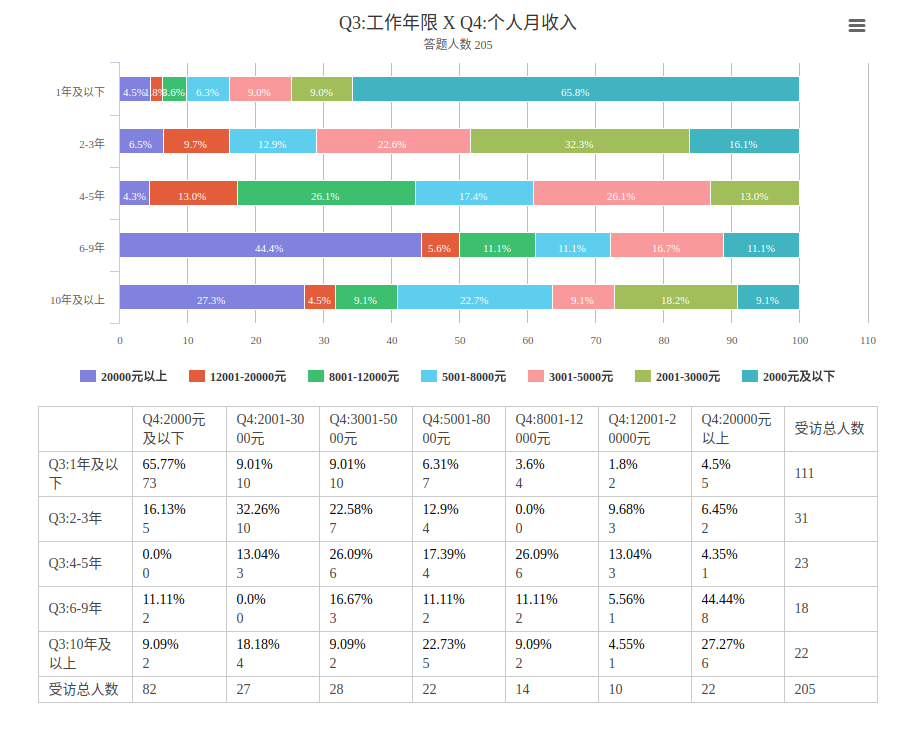
利用社交软件发放调查问卷，收集数据信息，两天半共获得调查问卷205份其中男性占比67.8%，女性占比32.2%。来自山东的调查问卷位39%



1. 进行数据清晰，数据无量纲化与归一化

对获得的数据进行初步处理，首先去除调查问卷附带的来源信息，包括IP地址，访问设备，时间等。然后将调查问卷中的吸引力要素进行无量纲化，将非常能够，比较能够，一般，不能够转化为4,3,2,1，然后将其归一化。同时将是否有吸引力问题的答案转化为1和0。

1. 去除部分异常数据



对于原始数据，部分数据存在明显的异常，但是因为数据量与维度都较小，子空间异常点检测算法与区域异常点检测算法都存在较大的误差，因此通过数据分析。设定人工规则筛选异常点，筛选掉工作十年以上且月收入在3000元以下，以及工作六年以上且月收入在2000元以下的点。然后通过多种分类算法进行前后对比，训练误差与测试误差均有所下降。

**5.2问题一模型的建立与求解**

**5.2.1特征选择**

传统的统计方法对于模型特征选取存在较强的主观性，但是机器学习算法对于特征的选取则更客观。本文使用包裹式特征选择对特征进行选择。

包裹式特征选择即选择一个基学习器，然后在此之上进行进行循环特征划分，然后将不同的特征划分后的数据集进行n折交差验证，然后按照评价函数得到在这些特征上最佳的特征划分方式，以及不同特征划分方式的评价分数。

交叉验证是用于防止模型过于复杂而引起的过拟合.有时亦称循环估计， 是一种统计学上将数据样本切割成较小子集的实用方法。于是可以先在一个子集上做分析， 而其它子集则用来做后续对此分析的确认及验证。 一开始的子集被称为训练集。而其它的子集则被称为验证集或测试集。交叉验证是一种评估统计分析、机器学习算法对独立于训练数据的数据集的泛化能力（generalize）。

而对于包裹式特征选择中基学习器的误差问题。设误差W为包裹式特征选择对于每类特征缺失时交叉验证得出的平均准确率与最佳类别特征选取的平均准确率的差值，若W小于整体数据集在同样次数的交叉验证中的最大准确率与最小准确率的差值，则判定为正常。在特征选取之后，得到的结果证明，之前建立的特征均在合理的误差范围之内。

**5.2.2对数据进行Logistics逻辑回归分析**

逻辑回归（Logistic regression ），即逻辑模型是离散选择法模型之一，一种广义的线性回归分析模型，属于多重变量分析范畴，是社会学、生物统计学、临床、数量心理学、计量经济学、市场营销等统计实证分析的常用方法。例如，探讨购买意向的因素，并根据购买意向因素预测是否购买等。以购车购买意向分析为例，学则两组人群，一组为购买组，一组为非购买组，两组人群必定有不同的生活标准和生活方式等，因此因变量为是否购车，值为“是”或“否”，自变量就可以很多了，如年例、性别、收入、生活标准等。自变量可以是连续或分类的。然后通过逻辑回归分析，得到自变量的权重，从而可以大致了解到那些因素是购车意向因素。同时根据权重可以预测一个人是购车的可能性。

Logistics回归分析是比较常用的变广义线性回归分析模型，其自变量既可以是连续的也可以是分类的。通过Logistics回归分析可以得到自变量的权重，从而大致了解到底哪些因素是影响结果的原因。

模型建立如下：

  **（公式 5-1）**

**5.2.3 模型分析**

对于人才整体而言对于济南子女教育，产业发展，户籍政策，基础设施建设，休闲娱乐都存在着积极影响，但是济南的企业条件，住房条件则对其存在负影响,。济南的产业发展每提高百分之一，那么其对人才的吸引力就会增加0.603%，子女教育每提高百分之一，那么对人才的吸引力就提高0.635%，等。诚然相比于北上广深等一线城市，济南的子女教育问题与落户更好解决，但是济南的房价相对于其收入水平而言也确实存在着日益不成正比的问题。

**5.2.4 放管服政策对于济南市人才吸引力提高的影响。**

济南市人才吸引力模型中，放管服政策主要对应于产业发展。其中此吸引力要素的权重为0.603。也就是说济南的产业发展每提高百分之一，那么其对人才的吸引力就会增加0.603%。而根据数据调查，济南市2017年第一季度与2016年第一季度公共财政收入增加了31.7%，企业注册数量比过去数年都多。因此我们可以判定济南市的放管服政策对济南的产业发展产生了必然的有益影响。其中济南市公共财政收入与企业发展基本挂钩，以此为量化标准，则可以判定在2016到2017这一年度，济南市放管服政策对济南市人才吸引力提升了19.115%

**5.2问题问模型的建立与求解**

**5.2.1 济南市人才需类型分析**

根据济南市人才引进计划，济南市2016年共引进人才1474位，其中对于博士197位，硕士197位，本科391位，学位不限689位，总人数共1474。根据济南市数年来的人才引进计划，可以将济南的人才需求划分为大致三种类型，一，科研教育类，包括科学家，大学教授，学术带头人等；二，技术研发类，包括机械与电子产业的工程师，设计师，技术顾问等，三，管理销售类，高级管理人才与销售人才，此类人才在数据分布上具有较强的一致性。

**5.2.2 济南市人才吸引力按类型建模**

根据济南市人才类型进行Logistics回归分析：

1. 科研教育类模型



**（公式 5-2）**

1. 管理销售人才模型

 **（公式 5-3）**

（3）技术研发人才模型

 **（公式 5-4）**

1. 其他类型模型

 **（公式 5-5）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 人才整体 | 技术研发类人才 | 管理销售类 | 科研医疗类 |
| 基础设施建设 | 0.471 | 0.549 | 0.351 | -0.287 |
| 生态环境 | -0.123 | -0.309 | 0.394 | -0.362 |
| 治安环境 | 0.105 | -0.109 | -0.082 | 0.626 |
| 产业发展 | 0.602 | -0.056 | 1.128 | 0.536 |
| 企业条件 | -0.787 | -0.310 | -0.929 | 0.171 |
| 职业发展 | 0.253 | 0.524 | -0.778 | 0.750 |
| 收入水平 | 0.041 | -0.358 | 0.437 | -0.335 |
| 住房条件 | -0.639 | 0.099 | -0.942 | -0.483 |
| 户籍政策 | 0.451 | -0.182 | 0.676 | 1.391 |
| 配偶就业 | -0.339 | -0.008 | -0.287 | -0.794 |
| 子女教育 | 0.635 | 0.578 | 0.352 | 0.265 |
| 生活成本 | -0.049 | 0.106 | 0.316 | -0.240 |
| 交通状态 | -0.184 | -0.690 | -0.339 | -0.246 |
| 休闲娱乐 | 0.376 | 0.845 | 0.271 | -0.662 |

**（表一）**

**5.2.3 济南市与同类型城市天津进行对比，并提出相关政策**

根据我们建立的模型与《天津市人才吸引力影响要素评价\_基于随机森林法的检验分析》中的天津市人才吸引力模型进行纵向对比可以得出天津与济南在地区人才吸引力上的差距，其中天津市在住房条件，生态环境，产业发展三个方面对人才存在较显著的吸引力影响，其中济南的产业发展，职业发展与住房条件在三种类型的人才吸引力上都存在较大劣势，但是济南户籍政策，子女教育，休闲娱乐，生态环境四个方面基本都存在优势，而对于科研医疗类人才而言，济南虽然不在休闲娱乐方面存在优势，但是在企业条件上存在一定优势；对于研发开发类人才而言，济南则在交通状态方面使其不满，但是在休闲娱乐方面的表现则极为突出；对于管理销售人才而言，济南拥有更引人注目的收入但是在住房条件与企业条件方则不招人喜欢。同时非常引人瞩目的是济南市人才引入薪资水平，博士仅为20.1万/年，而天津则为23.0万/年；济南的硕士仅为11.0万/年，而天津位15.0万/年。而不=不限制学位的人才引进中，济南平均薪资为15.5万/年，而天津为15.7万/年，可以看出相较于天津，济南的收入确实存在不小的差异

通过对比我们提出相应的政策为：

1. 对济南市引进人才进行收入补贴，减少收入水平对济南市人才吸引力的负面作用
2. 对济南市引进人才进行购房补贴，提高济南市人才受住房条件的负面影响
3. 扩大济南市在家庭保障维度与生活质量维度的优势，进一步提高济南对不同类型人才的吸引力

**5.3 问题三的建模与解决**

**5.3.1 济南市高新区的经济技术发展特点和相关人才政策分析**

济南市高新区是国务院批准设立的国家高新技术产业开发区，其以汽车、机械、轻化、食品、建材为主导产业的数十种产品在国内外享有盛名。并且其在本区内拥有多个商业区，现在的区域建设中，人才需求则以科研，金融，管理，电子信息等领域为主。

近些年来济南市高新区的的人才相关政策，重点在于鼓励人才创新创业，并非社会做出贡献的人才根据标准分档发放奖励金，同时对科研人才进行更高的生活补贴与住房租房补贴。同时对女性人才的子女教育问题进行安排和相应资金补助。

**5.3.2 根据人才的动态发展进行Logistics回归分析**

人才在不同的时期发展时期存在不同的需求，因此对于不同发展时期的人才而言，人才吸引力要素对其影响力与影响方式是不同的。对我们搜集和处理的数据按照工作年限进行划分，1年及以下为短期，2年到5年为长期，六年及以上为长期。

建立模型如下：

1. 短期模型

 **（公式 5-5）**

1. 中期模型

 **（公式 5-6）**

1. 长期模型

 **（公式 5-7）**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 短期 | 中期 | 长期 |
| 基础设施建设 | 0.302 | 0.071 | 0.524 |
| 生态环境 | -0.361 | 0.186 | 0.212 |
| 治安环境 | 0.363 | -0.588 | 0.123 |
| 产业发展 | 0.596 | -0.069 | 0.343 |
| 企业条件 | -0.927 | -0.183 | -0.167 |
| 职业发展 | 0.367 | 0.144 | -0.299 |
| 收入水平 | -0.066 | -0.160 | 0.493 |
| 住房条件 | -0.842 | 0.047 | 0.078 |
| 户籍政策 | 0.678 | 0.019 | 0.165 |
| 配偶就业 | -0.252 | -0.191 | -0.807 |
| 子女教育 | 0.579 | 0.723 | -0.136 |
| 生活成本 | -0.260 | 0.355 | 0.065 |
| 交通状态 | -0.436 | 0.211 | 0.359 |
| 休闲娱乐 | 0.867 | 0.229 | -0.050 |

**（表二）**

**5.3.3 通过模型提出政策建议**

从模型中我们可以得到人才在不同时期不同的发展需求。

1. 人才在工作初期更在意休闲娱乐，户籍政策与产业发展，子女教育，职业发展五个因素，因此对于刚工作的人才应当降低其迁移成本，提高其生活质量与家庭保障，并适当的提供租房补贴，同时应当进一步深化放管服改革，促进产业发展，提高相关方面对人才的吸引能力。而对于子女教育，生活成本，休闲娱乐，交通状态与生态环境五个维度，因此应当更好的解决其子女教育问题，并且进一步改善济南市高新区的交通状况，并为人才生活质量提供保障与支持。
2. 对于工作时间为中期的人而言，其更看重子女教育与生活成本，交通状态，休闲娱乐这三个生活质量维度的内容，同时事业维度的影响降低。因此对于处于这一发展时期的人才，应当注重其生活质量的保障，并同时提高其事业发展的期望程度，提高此类维度的影响力。因此可以对处于这一时期的人才提供部分旅游，假期补助。并适当向下延伸高新区原有的按工作职位奖励金分档，让更多该时期的人才获得奖励。

（3）对于工作时间已经较长的人才而言，其更注重基础设施建设，收入水平，交通状态以及产业发展。因此对于已经有充足工作经验的人才，应当从公共状况与产业发展角度入手，同时注重生态环境建设，例如对一环一湖的建设，在放管服改革的同时，改善公民的一些政务与手续办理。更好的使其感受济南生活的便利与舒适。

1. 模型的评价

**6.1模型优点**

（1）建立了包裹式特征分析基学习器判断标准，避免了以往的主观判断不准确问题。

（2）出色的调查问卷设计为人才分类型分析与动态分析提供了数据支持。

**6.2模型缺点**

1. 数据量较小，调查不完全符合随机抽样调查标准。
2. 放管服量化体系较为粗糙，忽略了较多的影响因素。
3. 数据量小，后期分类型模型出现无法进行卡方检验的问题。

1. 模型的改进

（1）根据随机抽样调查进行更大规模的数据调查。

（2）建立一个专业放管服改革量化评价体系。

1. 参考文献

[1]魏浩、王宸、毛日昇：《国际间人才流动及其影响因素的实证分析》，《管理世界》,2012年第１期

[2]张炜 :天津市人才吸引力影响要素评价\_基于随机森林法的检验分析,天津市社会科学界学术年会-2017

1. 附录

Python on jupyter notebook部分源代码完整代码与数据详见支撑材料：

第一问\_数据预处理：

import pandas as pd

path = "//home/font/Data/MM/JiNan/"

ppd\_list = 'jndata.csv'

df\_lcis = pd.read\_csv((path+ppd\_list))

len(df\_lcis)

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['答卷编号'],axis=1)

df\_lcis\_drop\_copy['sex'] = 1

df\_lcis\_drop\_copy['sex'].loc[df\_lcis\_drop\_copy['Q1\_性别'] == '男'] = 2

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q1\_性别'],axis=1)

df\_lcis\_drop\_copy['province'] = 2

df\_lcis\_drop\_copy['province'].loc[df\_lcis\_drop\_copy['Q2\_您所在的城市是\_省份'] == '山东省'] = 1

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q2\_您所在的城市是\_省份'],axis=1)

df\_lcis\_drop\_copy['city'] = 2

df\_lcis\_drop\_copy['city'].loc[df\_lcis\_drop\_copy['Q2\_您所在的城市是\_城市'] == '济南市'] = 1

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q2\_您所在的城市是\_城市','Q2\_您所在的城市是\_区/县'],axis=1)

df\_lcis\_drop\_copy['工作年限'] = 1

df\_lcis\_drop\_copy['工作年限'].loc[df\_lcis\_drop\_copy['Q3\_工作年限'] == '2-3年'] = 2

df\_lcis\_drop\_copy['工作年限'].loc[df\_lcis\_drop\_copy['Q3\_工作年限'] == '4-5年'] = 3

df\_lcis\_drop\_copy['工作年限'].loc[df\_lcis\_drop\_copy['Q3\_工作年限'] == '6-9年'] = 4

df\_lcis\_drop\_copy['工作年限'].loc[df\_lcis\_drop\_copy['Q3\_工作年限'] == '10年及以上'] = 5

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q3\_工作年限'],axis=1)

# df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q4\_个人月收入'],axis=1)

df\_lcis\_drop\_copy['基础设施建设'] = 1

df\_lcis\_drop\_copy['基础设施建设'].loc[df\_lcis\_drop\_copy['Q5\_基础设施建设'] == '一般'] = 2

df\_lcis\_drop\_copy['基础设施建设'].loc[df\_lcis\_drop\_copy['Q5\_基础设施建设'] == '比较能够'] = 3

df\_lcis\_drop\_copy['基础设施建设'].loc[df\_lcis\_drop\_copy['Q5\_基础设施建设'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q5\_基础设施建设'],axis=1)

df\_lcis\_drop\_copy['生态环境'] = 1

df\_lcis\_drop\_copy['生态环境'].loc[df\_lcis\_drop\_copy['Q6\_生态环境(包括城市绿化,城市景观生态文明等)'] == '一般'] = 2

df\_lcis\_drop\_copy['生态环境'].loc[df\_lcis\_drop\_copy['Q6\_生态环境(包括城市绿化,城市景观生态文明等)'] == '比较能够'] = 3

df\_lcis\_drop\_copy['生态环境'].loc[df\_lcis\_drop\_copy['Q6\_生态环境(包括城市绿化,城市景观生态文明等)'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q6\_生态环境(包括城市绿化,城市景观生态文明等)'],axis=1)

df\_lcis\_drop\_copy['治安环境'] = 1

df\_lcis\_drop\_copy['治安环境'].loc[df\_lcis\_drop\_copy['Q7\_治安环境'] == '一般'] = 2

df\_lcis\_drop\_copy['治安环境'].loc[df\_lcis\_drop\_copy['Q7\_治安环境'] == '比较能够'] = 3

df\_lcis\_drop\_copy['治安环境'].loc[df\_lcis\_drop\_copy['Q7\_治安环境'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q7\_治安环境'],axis=1)

df\_lcis\_drop\_copy['产业发展'] = 1

df\_lcis\_drop\_copy['产业发展'].loc[df\_lcis\_drop\_copy['Q8\_产业发展'] == '一般'] = 2

df\_lcis\_drop\_copy['产业发展'].loc[df\_lcis\_drop\_copy['Q8\_产业发展'] == '比较能够'] = 3

df\_lcis\_drop\_copy['产业发展'].loc[df\_lcis\_drop\_copy['Q8\_产业发展'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q8\_产业发展'],axis=1)

df\_lcis\_drop\_copy['企业条件'] = 1

df\_lcis\_drop\_copy['企业条件'].loc[df\_lcis\_drop\_copy['Q9\_企业条件'] == '一般'] = 2

df\_lcis\_drop\_copy['企业条件'].loc[df\_lcis\_drop\_copy['Q9\_企业条件'] == '比较能够'] = 3

df\_lcis\_drop\_copy['企业条件'].loc[df\_lcis\_drop\_copy['Q9\_企业条件'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q9\_企业条件'],axis=1)

df\_lcis\_drop\_copy['职业发展'] = 1

df\_lcis\_drop\_copy['职业发展'].loc[df\_lcis\_drop\_copy['Q10\_职业发展'] == '一般'] = 2

df\_lcis\_drop\_copy['职业发展'].loc[df\_lcis\_drop\_copy['Q10\_职业发展'] == '比较能够'] = 3

df\_lcis\_drop\_copy['职业发展'].loc[df\_lcis\_drop\_copy['Q10\_职业发展'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q10\_职业发展'],axis=1)

df\_lcis\_drop\_copy['收入水平'] = 1

df\_lcis\_drop\_copy['收入水平'].loc[df\_lcis\_drop\_copy['Q11\_收入水平'] == '一般'] = 2

df\_lcis\_drop\_copy['收入水平'].loc[df\_lcis\_drop\_copy['Q11\_收入水平'] == '比较能够'] = 3

df\_lcis\_drop\_copy['收入水平'].loc[df\_lcis\_drop\_copy['Q11\_收入水平'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q11\_收入水平'],axis=1)

df\_lcis\_drop\_copy['住房条件'] = 1

df\_lcis\_drop\_copy['住房条件'].loc[df\_lcis\_drop\_copy['Q12\_住房条件'] == '一般'] = 2

df\_lcis\_drop\_copy['住房条件'].loc[df\_lcis\_drop\_copy['Q12\_住房条件'] == '比较能够'] = 3

df\_lcis\_drop\_copy['住房条件'].loc[df\_lcis\_drop\_copy['Q12\_住房条件'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q12\_住房条件'],axis=1)

df\_lcis\_drop\_copy['户籍政策'] = 1

df\_lcis\_drop\_copy['户籍政策'].loc[df\_lcis\_drop\_copy['Q13\_户籍政策'] == '一般'] = 2

df\_lcis\_drop\_copy['户籍政策'].loc[df\_lcis\_drop\_copy['Q13\_户籍政策'] == '比较能够'] = 3

df\_lcis\_drop\_copy['户籍政策'].loc[df\_lcis\_drop\_copy['Q13\_户籍政策'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q13\_户籍政策'],axis=1)

df\_lcis\_drop\_copy['配偶就业'] = 1

df\_lcis\_drop\_copy['配偶就业'].loc[df\_lcis\_drop\_copy['Q14\_配偶就业'] == '一般'] = 2

df\_lcis\_drop\_copy['配偶就业'].loc[df\_lcis\_drop\_copy['Q14\_配偶就业'] == '比较能够'] = 3

df\_lcis\_drop\_copy['配偶就业'].loc[df\_lcis\_drop\_copy['Q14\_配偶就业'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q14\_配偶就业'],axis=1)

df\_lcis\_drop\_copy['子女教育'] = 1

df\_lcis\_drop\_copy['子女教育'].loc[df\_lcis\_drop\_copy['Q15\_子女教育'] == '一般'] = 2

df\_lcis\_drop\_copy['子女教育'].loc[df\_lcis\_drop\_copy['Q15\_子女教育'] == '比较能够'] = 3

df\_lcis\_drop\_copy['子女教育'].loc[df\_lcis\_drop\_copy['Q15\_子女教育'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q15\_子女教育'],axis=1)

df\_lcis\_drop\_copy['生活成本'] = 1

df\_lcis\_drop\_copy['生活成本'].loc[df\_lcis\_drop\_copy['Q16\_生活成本'] == '一般'] = 2

df\_lcis\_drop\_copy['生活成本'].loc[df\_lcis\_drop\_copy['Q16\_生活成本'] == '比较能够'] = 3

df\_lcis\_drop\_copy['生活成本'].loc[df\_lcis\_drop\_copy['Q16\_生活成本'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q16\_生活成本'],axis=1)

df\_lcis\_drop\_copy['交通状态'] = 1

df\_lcis\_drop\_copy['交通状态'].loc[df\_lcis\_drop\_copy['Q17\_交通状态'] == '一般'] = 2

df\_lcis\_drop\_copy['交通状态'].loc[df\_lcis\_drop\_copy['Q17\_交通状态'] == '比较能够'] = 3

df\_lcis\_drop\_copy['交通状态'].loc[df\_lcis\_drop\_copy['Q17\_交通状态'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q17\_交通状态'],axis=1)

df\_lcis\_drop\_copy['休闲娱乐'] = 1

df\_lcis\_drop\_copy['休闲娱乐'].loc[df\_lcis\_drop\_copy['Q18\_休闲娱乐'] == '一般'] = 2

df\_lcis\_drop\_copy['休闲娱乐'].loc[df\_lcis\_drop\_copy['Q18\_休闲娱乐'] == '比较能够'] = 3

df\_lcis\_drop\_copy['休闲娱乐'].loc[df\_lcis\_drop\_copy['Q18\_休闲娱乐'] == '非常能够'] = 4

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q18\_休闲娱乐'],axis=1)

# df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么'],axis=1)

df\_lcis\_drop\_copy['score'] = 1

df\_lcis\_drop\_copy['score'].loc[df\_lcis\_drop\_copy['Q20\_请问你觉得济南这座城市对您是否有吸引力呢'] == '不作考虑'] = 2

df\_lcis\_drop\_copy['score'].loc[df\_lcis\_drop\_copy['Q20\_请问你觉得济南这座城市对您是否有吸引力呢'] == '一般'] = 3

df\_lcis\_drop\_copy['score'].loc[df\_lcis\_drop\_copy['Q20\_请问你觉得济南这座城市对您是否有吸引力呢'] == '有较强的吸引力'] = 4

df\_lcis\_drop\_copy['score'].loc[df\_lcis\_drop\_copy['Q20\_请问你觉得济南这座城市对您是否有吸引力呢'] == '不二选择'] = 5

df\_lcis\_drop\_copy = df\_lcis\_drop\_copy.drop(['Q20\_请问你觉得济南这座城市对您是否有吸引力呢'],axis=1)

df\_lcis\_drop\_copy.to\_csv((path+"finalData.csv"),index=False)

第一问\_模型调优:

import pandas as pd

path = "//home/font/Data/MM/JiNan/"

ppd\_list = 'finalData.csv'

df\_lcis = pd.read\_csv((path+ppd\_list))

df\_lcis2 = df\_lcis.drop(['Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么'],axis=1)

df\_lcis2['Score'] = 0

df\_lcis2['Score'].loc[df\_lcis['score'] == 3] = 1

df\_lcis2['Score'].loc[df\_lcis['score'] == 4] = 1

df\_lcis2['Score'].loc[df\_lcis['score'] == 5] = 1

df\_lcis2['个人月收入'] = 0

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '2001-3000元'] = 1

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '3001-5000元'] = 2

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '5001-8000元'] = 3

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '8001-12000元'] = 4

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '12001-20000元'] = 5

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '20000元以上'] = 6

# df\_lcis2 = df\_lcis2[(df\_lcis2['个人月收入'].isin([1,2,3,4,5,6]))]

# df\_lcis2 = df\_lcis2[(df\_lcis2['province'].isin([1]))]

df\_lcis2 = df\_lcis2.drop(['score','Q4\_个人月收入','个人月收入','工作年限','sex','province','city'],axis=1)

columns = df\_lcis2.columns

print(list(columns))

import numpy as np

from sklearn import preprocessing,model\_selection,linear\_model

X\_train = np.array(df\_lcis2.drop(["Score"], axis=1).values)

y\_train = np.array(df\_lcis2["Score"].values)

X\_train = preprocessing.scale(X\_train)

cv\_split = model\_selection.ShuffleSplit(n\_splits=10, test\_size=.3, train\_size=.7,random\_state=0)

cv\_results = model\_selection.cross\_validate(linear\_model.LogisticRegressionCV(), X\_train, y\_train, cv=cv\_split)

print(cv\_results['train\_score'].mean())

print(cv\_results['test\_score'].mean())

def Mychisquare(\*data):

y\_test,y\_pre=data

import scipy

chisq, p=scipy.stats.chisquare(y\_test,y\_pre)

print(chisq, p)

# print("chisq----> %d,p值----> %d"%(chisq,p))

def LogisticRegression\_multinomial(\*data):

X\_train,X\_test,y\_train,y\_test=data

regr = linear\_model.LogisticRegression()

regr.fit(X\_train, y\_train)

y\_pre = regr.predict(X\_test)

Mychisquare(y\_test,y\_pre)

print('Score: %.4f' % regr.score(X\_test, y\_test))

print('Coefficients:%s, intercept %s'%(regr.coef\_,regr.intercept\_))

showCoef(regr.coef\_)

def showCoef(listCoef):

print("下面是各项指标的权重")

for i in range(len(listCoef[0])):

print(" ",i," ",df\_lcis2.columns[i]," : ",listCoef[0][i])

print("\nREFCV属性筛选---->all")

f = listCoef[0][3]\*31.7

print("\n以公共财政收入同比增长为标准,济南市高新区放管服改革对人才的吸引力增加了百分之 %.3f"%f)

LogisticRegression\_multinomial(X\_train,X\_train,y\_train,y\_train)

第二问\_不同类型人才不同城市之间区别差异：

import pandas as pd

path = "//home/font/Data/MM/JiNan/"

ppd\_list = 'finalData.csv'

df\_lcis = pd.read\_csv((path+ppd\_list))

# df\_lcis2 = df\_lcis.drop(['Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么'],axis=1)

list(set(df\_lcis['Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么']))

df\_lcis2 = df\_lcis

df\_lcis2['Score'] = 0

df\_lcis2['Score'].loc[df\_lcis['score'] == 3] = 1

df\_lcis2['Score'].loc[df\_lcis['score'] == 4] = 1

df\_lcis2['Score'].loc[df\_lcis['score'] == 5] = 1

df\_lcis2['个人月收入'] = 0

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '2001-3000元'] = 1

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '3001-5000元'] = 2

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '5001-8000元'] = 3

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '8001-12000元'] = 4

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '12001-20000元'] = 5

df\_lcis2['个人月收入'].loc[df\_lcis['Q4\_个人月收入'] == '20000元以上'] = 6

df\_lcis2 = df\_lcis2.drop(['score','Q4\_个人月收入','个人月收入','工作年限','sex','province','city'],axis=1)

columns = df\_lcis2.columns

def Mychisquare(\*data):

y\_test,y\_pre=data

import scipy

chisq, p=scipy.stats.chisquare(y\_test,y\_pre)

print(" chisq----> %d,p值----> %d"%(chisq,p))

def LogisticRegression\_multinomial(\*data):

from sklearn import linear\_model

X\_train,X\_test,y\_train,y\_test,columns=data

regr = linear\_model.LogisticRegression()

regr.fit(X\_train, y\_train)

y\_pre = regr.predict(X\_test)

# Mychisquare(y\_test,y\_pre)

print('\n成绩为:\n Score: %.4f' % regr.score(X\_test, y\_test))

print('\n权重与常量为:\n Coefficients:%s, \n intercept %s'%(regr.coef\_,regr.intercept\_))

showCoef(regr.coef\_,columns)

def showCoef(listCoef,columns):

print("\n下面是各项指标的权重")

for i in range(len(listCoef[0])):

print(" ",i," ",columns[i]," : ",listCoef[0][i])

def testModel(list\_Q19):

print("\n对于","与".join(list\_Q19),"专业的人建模.结果如下:\n")

this\_df\_lcis = df\_lcis2[(df\_lcis2['Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么'].isin(list\_Q19))]

this\_df\_lcis = this\_df\_lcis.drop(['Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么'],axis=1)

import numpy as np

from sklearn import preprocessing

this\_X\_train = np.array(this\_df\_lcis.drop(["Score"], axis=1).values)

this\_y\_train = np.array(this\_df\_lcis["Score"].values)

LogisticRegression\_multinomial(this\_X\_train,this\_X\_train,this\_y\_train,this\_y\_train,this\_df\_lcis.columns)

testModel(["技术研发人才","技工人才"])

testModel(['服务售后', '金融从业者', '销售人才', '管理人员'])

testModel(['医疗', '教育科研'])

testModel(['其他'])

第三问\_根据人才发展轨迹进行人才吸引力建模：

import pandas as pd

path = "//home/font/Data/MM/JiNan/"

ppd\_list = 'finalData.csv'

df\_lcis = pd.read\_csv((path+ppd\_list))

ppd\_list\_old = 'jndata.csv'

df\_lcis\_old = pd.read\_csv((path+ppd\_list\_old))

df\_lcis2 = df\_lcis

df\_lcis2['Score'] = 0

df\_lcis2['Score'].loc[df\_lcis['score'] == 2] = 0

df\_lcis2['Score'].loc[df\_lcis['score'] == 3] = 1

df\_lcis2['Score'].loc[df\_lcis['score'] == 4] = 1

df\_lcis2['Score'].loc[df\_lcis['score'] == 5] = 1

df\_lcis2['工作年限'] = df\_lcis\_old['Q3\_工作年限']

list\_Q19 = ["金融从业者","技术研发人才","技工人才","管理人员","教育科研"]

this\_df\_lcis = df\_lcis2[(df\_lcis2['Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么'].isin(list\_Q19))]

df\_lcis2 = df\_lcis2.drop(['score','Q4\_个人月收入','sex','province','city','Q19\_请问您的主要职业或者毕业之后将要成为的主要职业是什么'],axis=1)

columns = df\_lcis2.columns

print(columns)

print(list(set(df\_lcis2["工作年限"])))

def Mychisquare(\*data):

y\_test,y\_pre=data

import scipy

chisq, p=scipy.stats.chisquare(y\_test,y\_pre)

print(" chisq----> %d,p值----> %d"%(chisq,p))

def LogisticRegression\_multinomial(\*data):

from sklearn import linear\_model

X\_train,X\_test,y\_train,y\_test,columns=data

regr = linear\_model.LogisticRegression()

# print(y\_train)

regr.fit(X\_train, y\_train)

y\_pre = regr.predict(X\_test)

# Mychisquare(y\_test,y\_pre)

# print('\n成绩为:\n Score: %.4f' % regr.score(X\_test, y\_test))

# print('\n权重与常量为:\n Coefficients:%s, \n intercept %s'%(regr.coef\_,regr.intercept\_))

print("常量 :",regr.intercept\_)

showCoef(regr.coef\_,columns)

def showCoef(listCoef,columns):

print("\n下面是各项指标的权重")

for i in range(len(listCoef[0])):

print(" ",i," ",columns[i]," : ",listCoef[0][i])

def testModel(list\_Q19):

print("\n对于工作年限为","与".join(list\_Q19),"的人建模.结果如下:\n")

this\_df\_lcis = df\_lcis2[(df\_lcis2['工作年限'].isin(list\_Q19))]

this\_df\_lcis = this\_df\_lcis.drop(['工作年限'],axis=1)

import numpy as np

from sklearn import preprocessing

this\_X\_train = np.array(this\_df\_lcis.drop(["Score"], axis=1).values)

this\_y\_train = np.array(this\_df\_lcis["Score"].values)

LogisticRegression\_multinomial(this\_X\_train,this\_X\_train,this\_y\_train,this\_y\_train,this\_df\_lcis.columns)

for i in [ ['1年及以下'], ['2-3年','4-5年'], ['6-9年','10年及以上']]:

testModel(i)