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：安徽大学

论文题目：拖地带娃擦玻璃，卷王竟在我家里？
——基于大数据挖掘与机器学习的家政行业整体素质提升因素分析

参赛队员：何志成 刘博今 任柏潼

指导老师：贾婧

拖地带娃擦玻璃，卷王竟在我家里？

——基于大数据挖掘与机器学习的家政行业整体素质提升因素分析

目录

表格清单

插图清单

摘要

家政服务业即为家庭提供多种类服务的专门行业，在第三产业中占有重要地位。但近年来，由于人工智能家居产业的发展与客户对家政从业者的要求水平不断提高，家政行业仍面对较大问题。

本文从家政从业人员的角度出发，首先，通过网络爬虫爬取家政从业者相关数据，并对数据进行量化处理后展开分析。其次，对家政从业者的工作经历和培训评价进行词频分析和词云图制作，以此来预判客户需求的倾向性。接着进行因子分析，得到家政从业者的综合评分公式，以此对家政从业者进行综合评估。而后，为进一步研究家政从业者的专业水平与核心素养对其制定预期工资的影响，在因子分析的基础上，使用回归模型进行验证，发现模型拟合度较好，说明家政从业者专业水平与其工资联系密切。在模型构建方面，我们建立了 BP 神经网络，并采用 Levenberg-Marquardt 算法仿真，得出准确率为 90.0% 的模型；同时，在使用模型筛选器将机器模型筛选出最适用的袋装树模型后，通过对训练集进行训练、优化剪枝最终得到较为简洁且采样误差与交叉验证误差分别为 0.1334 和 0.1735 的决策树模型，以此判断家政从业者的个人特质是否能使其个人工资超过行业平均工资。最后通过代入训练和袋外数据测试构建出准确性约为 90.5%的随机森林模型，分析影响家政从业者预期工资的因素。通过模型正确率分析得出，随机森林模型预测准确性较高。

通过分析得出结论，一方面，家政从业者想使得预期工资达到平均工资水平，须提升个人专业水平，参与正规专业培训，丰富个人专业技能，提升市场竞争力；另一方面，由于家政行业规模扩大，行业要求更加严格，行业秩序更加规范，家政行业整体素质有较大提升。

关键词：家政从业者；因子分析；BP 神经网络；决策树；随机森林

一、绪论

（一）背景

伴随着我国经济的发展，家政行业逐渐成为服务业的新星，在解决就业压力与吸纳社会剩余劳动力等方面发挥着重要作用。近年来，由于人口老龄化加剧、二孩政策逐步推进与人们对美好生活的需要，市场对中高端家政服务员的需求不断增加，社会各方人士逐渐关注到家政行业的广阔前景。

（二）现状

家政行业已转变成为社会公共的服务性组织，主要由专业家政人员承担相关事务。家政行业已由从前的“人人可做”转变成为全方面、高标准的服务模式。这些要求家政从业者除了具备烹饪、护理等基本工作外，还需要具备更高的文化素质与专业水平。与从前相比，人们对家政从业者的要求更加严格。

人工智能家居产业的发展对低端家政行业产生冲击。随着科技的发展，扫地机器人、智能管家AI等产品不断推出，在方便人们生活之余，与低端家政服务业产生替代关系。家政从业者为了提升个人市场竞争力，减少被机器替代的可能性，需要不断自我提升专业技能。家政从业者除了体力劳动外，更多的是体力与智力相结合的复杂劳动。

疫情冲击下，家政行业从线下产业转向线下与线上相结合。后疫情时代，人们切实感受到来自互联网模式的价值，家政行业尝试“互联网+家政”服务模式，使客户实现足不出户网上选人。在大数据环境下，从业者信息更加公开透明，其同行间竞争压力增大。

（三）文献综述

国内学者对家政服务业问题的研究始于90年代中期，近年来，伴随着放开三孩政策及老龄化的加剧，家政服务业已成为人们日常生活中不可缺少的部分。学者们对家政服务业问题的研究也日益高涨，他们从经济学、社会学、教育学、

法学、心理学等方面对家政问题进行了研究。

关于家政服务业的发展现状。学术界普遍认为家政服务业成为解决下岗职工再就业和转移农村富余劳力的有效途径，这一行业具有可观的发展潜力和美好前景，但家政服务热的背后也有着较为突出的矛盾。李艳梅（2008）[1] 认为，行业供给与市场需求间矛盾较突出。客户需要自己所雇佣的家政员受过正规培训并提供全方面高质量服务，但家政员又普遍受教育程度偏低，家政从业者供给不能很好满足客户需求，成为阻碍家政服务业的瓶颈。曹华 [2] 认为，专业的教育体系构建在家政服务业中表现不足，在市场上也缺少为家政服务人员提供参考借鉴的专业教材。唐海秀 [3] 认为，我国家庭服务业仍处于初步发展阶段，行业的供需链衔接不畅、诚信度水平低，存在一定的安全隐患，整个行业的发展不够规范与发达。

对家政服务业的未来发展的对策建议。有关家政服务业未来发展的对策主要就政府、企业、个人三方面予以分析：孙学致（2020）[4] 认为家政服务是吸纳劳动力，并且有效扩大就业的重要渠道。家政服务企业发展趋势呈现品牌化特征，服务更加专业精细，推动家政行业实现产业化发展。谷素萍（2019）[5] 认为，要整合家政服务资源，加力度培育大家政人才，使人员服务质量得到提升。杨军剑 [6] 指出，按照高质量发展的要求，积极开展家政教育和家政培训行。张贝妹 [7] 认为，应着眼于“互联网+家政”、“人工智能”、“区块链”等发展新业态，不断增加家政服务业发展的内生动力。

以往研究为本研究奠定了基础、开拓了思维，也为人们了解家政服务业提供了宝贵的经验。但大部分学者是从宏观层面研究家政行业，站在企业、政府等角度思考并提出解决意见，从家政从业者微观个体角度来研究家政产业的情况则较少。家政服务业是“小切口，大民生”的体现。为促进家政服务业提质扩容，实现高质量发展，以微观视角出发，从家政从业人员的角度，展开研究

以弥补家政服务业存在的缺陷短板、优化内需促进双循环发展具有一定必要性。本项目在以往研究的基础上，从微观个体角度、以第三产业中家政从业者为切入点来研究该群体综合服务质量提升因素，运用文献分析、大数据分析、建立模型等方式，从宏观层面和微观层面出发提出切实可行的对策建议。

二、研究内容与方法

（一）研究内容

对于家政服务从业人员的研究，本文进行了线上搜集数据，内容涉及基本信息、就业情况、培训状况、求职意向、工作经历、客户评价六大方面；对于家政服务需求人员进行线上调研，内容涉及家政服务人员满意度评价、问题及建议两方面内容；对于家政服务中介机构进行线上调研，内容涉及基本信息、雇佣状况、服务需求、培训方法、对策前景等五个方面。本次研究具体内容涵盖工作 id、实名信息、求职意向、是否住家、期望工资、年龄、学历、生肖、婚姻、籍贯、身高、体重、经验、状态、做饭口味、会说语言、工作范围、特殊技能、拥有证书、培训记录、工作记录，老师评价等。

（二）研究方法

本文选用 Python 软件以网络爬虫方式来进行数据收集，以家政港作为平台爬取文本信息，了解家政从业人员基本信息、个人状况、客户评价等。在数据收集后，运用 Excel 进行数据量化处理，随后运用 Python 软件进行词频分析和词云图程序编写；运用 Stata 进行因子分析与回归分析；运用 Matlab2018b 软件对 BP 神经网络模型的数据进行归一化处理与仿真模拟；运用 DevC++与 Matlab 共同编程实现有关随机森林模型的模型分类筛选、决策树和随机森林主函数的编写以及随机森林程序的运行；运用 Mathtype 进行相关数学公式的编写；运用 Photoshop、Tableau、Visio2013 分别进行词云图图片处理、地图和因子分析树状图以及流程图神经网络图的绘制，以此来进行统计推断的估计、统计分组确定、综合指标的测算、统计模型的衡量等。

（三）研究流程

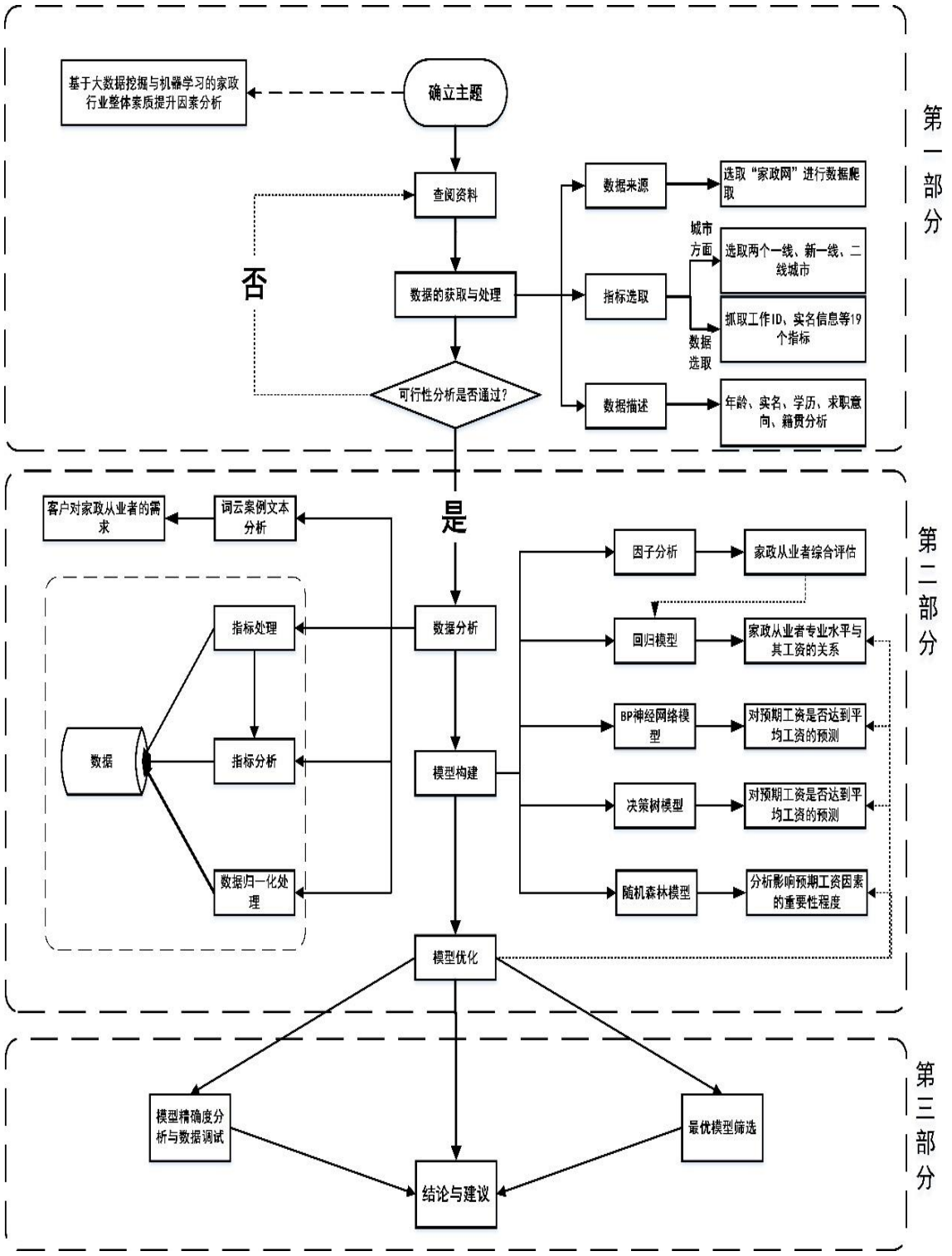


图 1　技术流程图

三、数据处理

（一）数据来源

在选取样本数据时，为保证数据信息数量充足、数据来源真实可靠，我们选取了“家政港”网站作为数据选取平台，并使用网络爬虫技术，获取了该网站截至 2021 年 5 月 11 日更新的六个城市其中 10800 名家政从业人员信息，随后将这些数据进行处理与分析。

本文主要进行相关数据查询的网站如下：

家政港：https://www.jiazhenggang.com/

国家统计局：http://www.stats.gov.cn/

职友集：https://www.jobui.com/

（二）指标选取

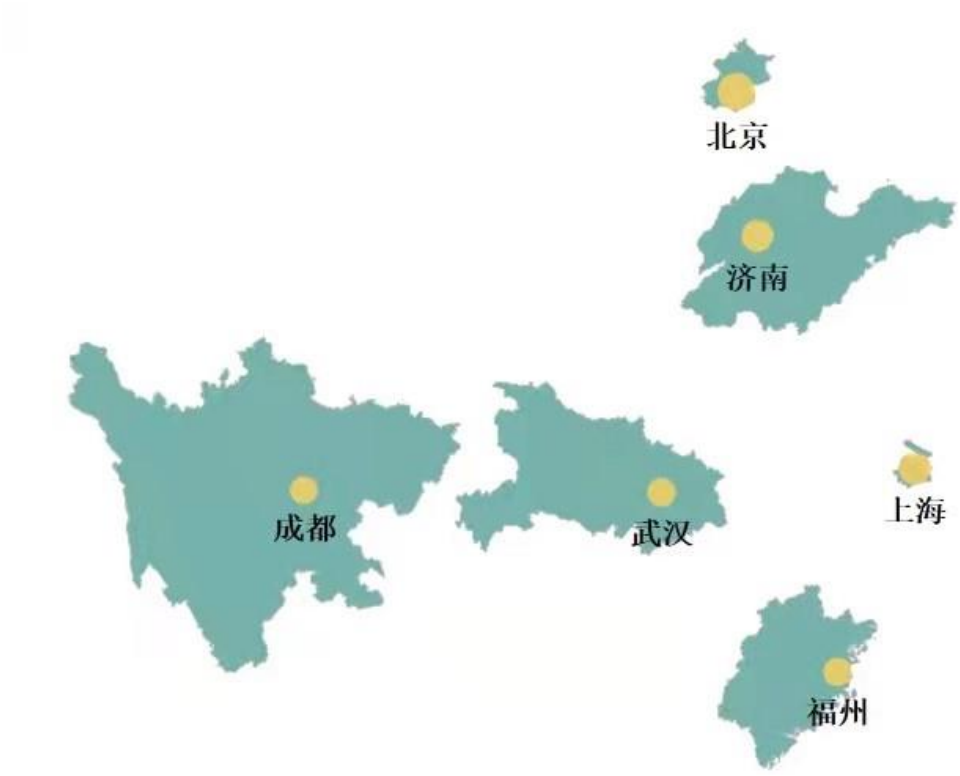


图 2　城市选取一览图

1.城市选取

在城市选取方面，为使样本更具代表性与普遍性，从而能够更好地反映总

体的特性。本文综合考量“2020 年全国一二三线城市划分标准”与城市地理区位因素差异，最终选取了处于一线城市的北京、上海，处于新一线城市的成都、武汉，与处在二线城市位置的济南、福州这六个城市进行研究，如图 2 所示。

2.指标选取

为了对家政从业人员有更加具体详细的了解，我们使用爬虫抓取了家政港网站中公开的从业人员详细信息，并归类整理成指标，具体如表 1 所示。

表 1　网络爬虫抓取信息表

调查平台	调查内容	抓取指标
家政港	调查城市：北京、上海、成都、武汉、福州、济南	工作 id
		实名信息
		是否住家
		期望工资
	调查板块：网上选人	平均期望工资
		年龄
		学历
		婚姻
	调查范围：各城市 1-30 页信息列表	籍贯
		身高
		体重
		经验
	调查对象：各人员主页详细信息	做饭口味
		会说语言
		工作范围
		特殊技能
	调查数量：10800 名从业人员信息	拥有证书
		求职意向
		老师评价

（三）数据描述

1.年龄分析

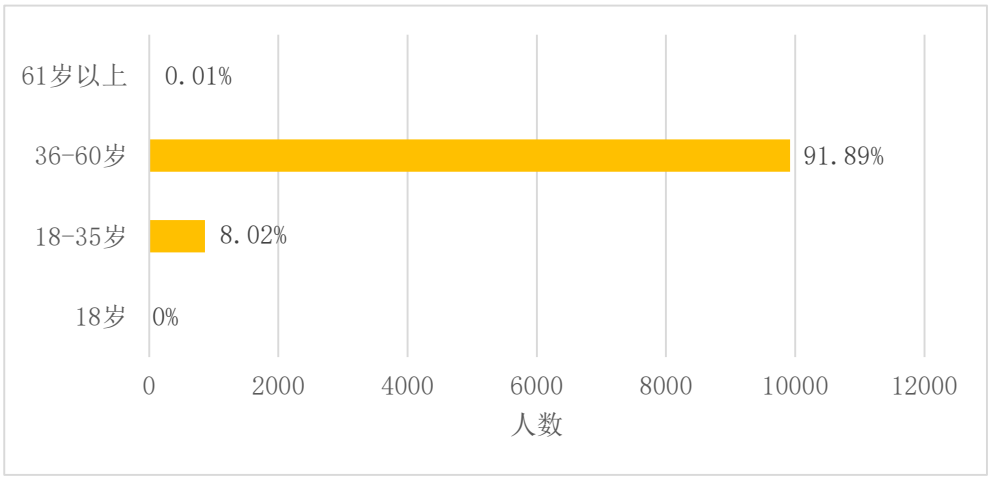


图 3　家政从业者年龄分布条形图

从图 3 中可以看出，家政从业者的年龄段主要集中在 36-60 岁之间，占比高达 91.89%，由于该行业从业者主要为女性，所以中年女性占该行业的主体地位。同时我们关注到，18-35 岁的青年人占比达到了 8.02%，说明该行业也有年轻人。60 岁以上与 18 岁以下从业者占比较少，合计不到 0.01%。这些数据表明青年家政人员供给存在缺口，未来可能出现服务断层。

2.实名分析

考虑到家政行业的规范性以及用人的可靠性与安全性，家政从业者的实名化是家政行业规范化道路上必须解决的问题。在对样本进行统计时发现，所选取的样本中，已实名人数达到 7878 人，占比达 72.94%，未实名人数有 2922 人，占比达 27.6%。由此观察出，虽然已实名人群占大部分，但未实名人群占比依旧很高。从业者通过实名可以加强客户对其的信任感，在市场竞争中被优先选择；同时政府也可以更好地管理家政产业人力资源，从而规范行业秩序，故家政人员的实名工作仍有待推进。

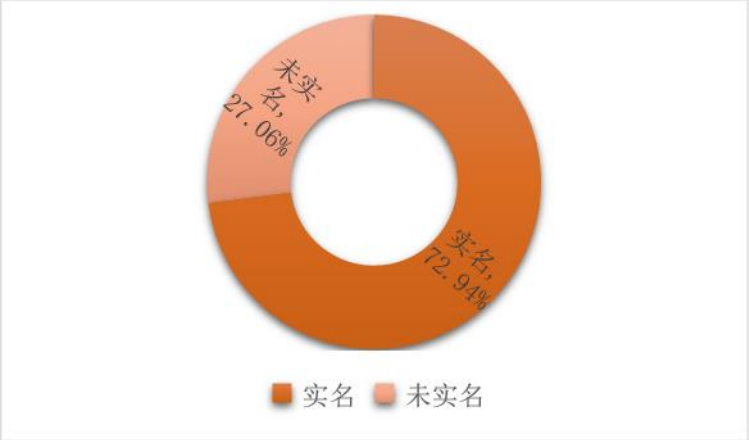


图 4　家政实名人员占比环形图

3.学历分析

家政从业者受教育水平一直广受关注，图 5 可见，家政从业人员学历为初中或中专占主体，达 51.99%；高中或大专学历占比较低；本科及以上学历从业者人数较少，占比仅为 1.77%；小学及以下学历占比最少，只有 1.37%。由此可见，家政从业者基本完成了 9 年义务教育，初中及以下学历占比不到 50%，与 2011 年 87%的数据相比有较大幅度降低 [8]。这说明近几年随着家政产业的进步与规范，从业者学历整体有所提高，对从业者学历要求也更加严格。

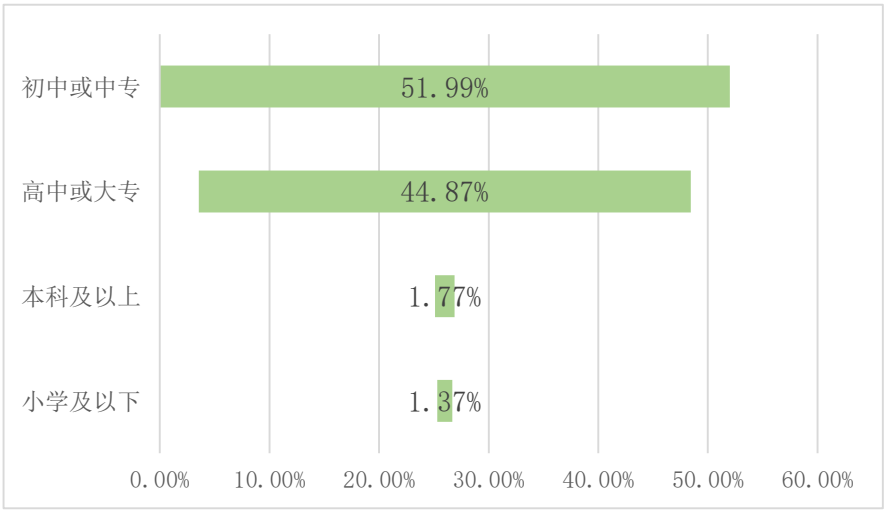


图 5　学历分布倒金字塔图

4.求职意向分析

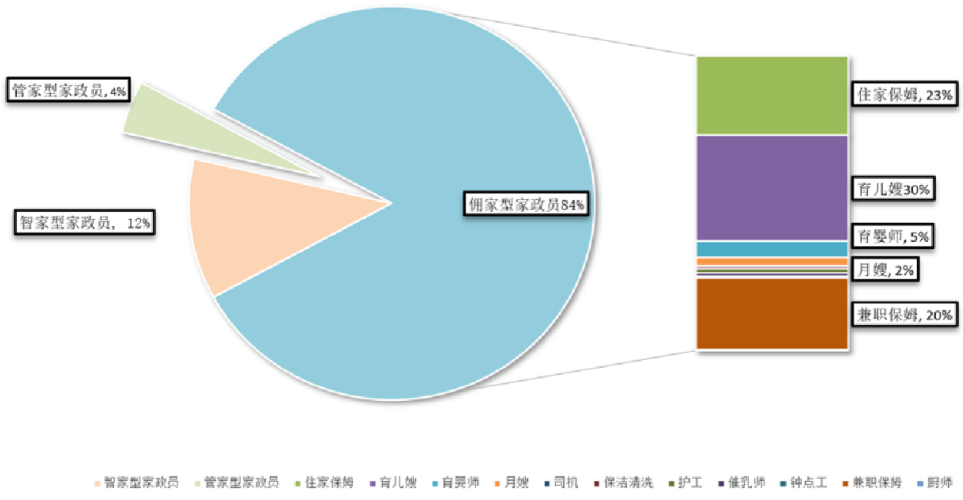
针对从业人员的就业意向，我们进行分类统计。由图 6 可知，求职意向占主要部分的是育儿嫂、住家保姆等职业，求职意向较少的为厨师、司机与幼儿家教等职业。对这些职业进行分类如表 2 所示。

表 2　求职意向分类表

分类	内容
智家型家政员	家教幼教、早教师、涉外保姆
管家型家政员	管家、高端保姆
佣家型家政员	住家保姆、育儿嫂、育婴师、司机、月嫂、保洁清洗、护工、催乳师、钟点工、兼职保姆、厨师

由图 6 知，佣家型家政员占比达到 84%，智家型家政员占比达到 12%，管家型家政员占比达到 4%。即 n 说明家政行业服务水平更高，提供了更高质量的服务。

图 6　家政从业人员求职意向分类饼图



5.籍贯分析

我们按城市对家政从业人员籍贯进行统计，如表 3、图 7 所示。

表 3　六城市各地家政从业人员来源地分布

	北京(%)	成都(%)	福州(%)	济南(%)	上海(%)	武汉(%)
北京	**0.02**	0.00	0.00	0.00	0.02	0.01
天津	0.00	0.01	0.01	0.00	0.00	0.03
上海	0.14	0.13	0.13	0.23	0.39	0.14
重庆	0.18	0.16	0.28	0.16	0.27	0.19
河北	0.22	0.05	0.08	0.11	0.09	0.09
山西	0.13	0.05	0.08	0.11	0.09	0.09
辽宁	0.19	0.12	0.25	0.16	0.18	0.18
吉林	0.14	0.05	0.20	0.11	0.10	0.06
黑龙江	0.56	0.57	0.26	0.50	0.24	0.52
江苏	**2.20**	**1.75**	**2.38**	**3.14**	**2.91**	**2.46**
浙江	0.58	0.17	0.68	0.49	0.50	0.44
安徽	**2.72**	**3.17**	**2.48**	**2.77**	**3.70**	**2.91**
福建	0.14	0.18	0.21	0.15	0.17	0.14
江西	0.64	0.89	0.65	0.71	0.81	0.72
山东	0.31	0.31	0.32	0.38	0.27	0.23
河南	**1.89**	**2.44**	**1.76**	**1.59**	**1.46**	**2.08**
湖北	**3.72**	**3.67**	**3.31**	**3.21**	**2.04**	**3.77**
湖南	0.40	0.56	0.60	0.47	0.49	0.45
广东	0.11	0.10	0.18	0.11	0.12	0.06
海南	0.03	0.02	0.02	0.01	0.01	0.03
四川	**1.10**	**1.31**	**1.31**	**1.22**	**1.61**	**1.08**
贵州	0.11	0.09	0.15	0.07	0.15	0.12
云南	0.08	0.04	0.11	0.02	0.14	0.06
陕西	0.53	0.51	0.73	0.52	0.48	0.46
甘肃	0.13	0.13	0.14	0.11	0.20	0.13
青海	0.02	0.00	0.01	0.00	0.00	0.00
台湾	0.01	0.00	0.00	0.01	0.01	0.00
内蒙古	0.00	0.00	0.02	0.02	0.03	0.02
广西	0.26	0.19	0.28	0.22	0.16	0.17
西藏	0.00	0.00	0.00	0.00	0.00	0.00
宁夏	0.06	0.00	0.00	0.04	0.01	0.00
新疆	0.02	0.02	0.04	0.02	0.01	0.02
香港	0.02	0.00	0.00	0.00	0.00	0.00
澳门	0.00	0.00	0.00	0.00	0.00	0.00
国外	**0.01**	0.00	0.00	0.00	**0.02**	**0.01**

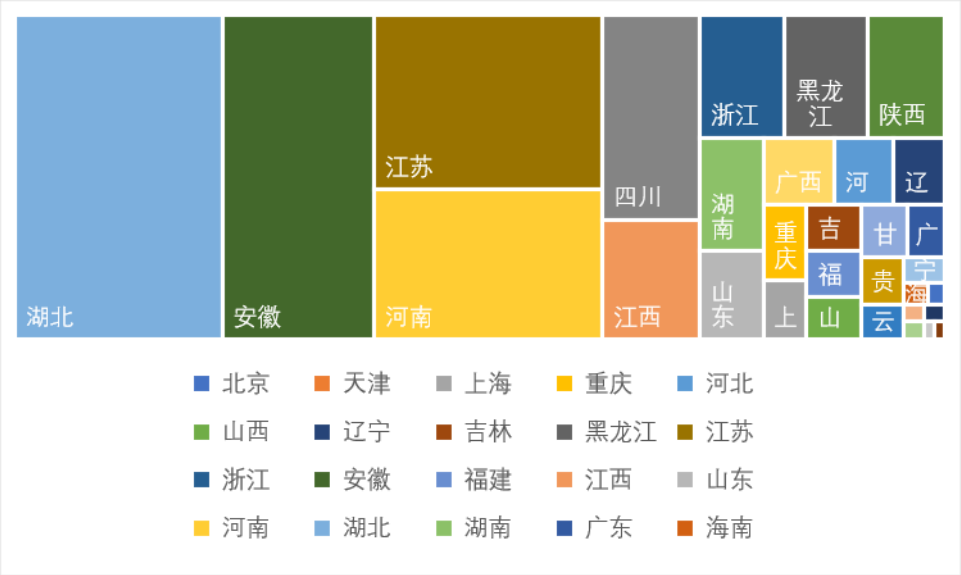


图 7　六城市各地家政从业人员籍贯分布树状统计图

由图表可见，家政从业人员户籍来自湖北、安徽、河南、江苏、四川等省份较多；经济发达城市如北京上海，偏远地区如新疆、西藏则较少。由此我们得出结论，在城市就业的家政从业者大多数为外来移民，可见外来劳动力侵占本地家政市场现象较为严重，其中以临近省份就业人员较多。该现象是地理区位因素、城市间差异与人力资源流动三方面共同作用的结果。同时我们还发现，在北京、上海、武汉三地，家政从业人员也包含外籍人员，这说明家政市场正在不断做大，吸引着更多样化主体的加入。

6.预期工资分析

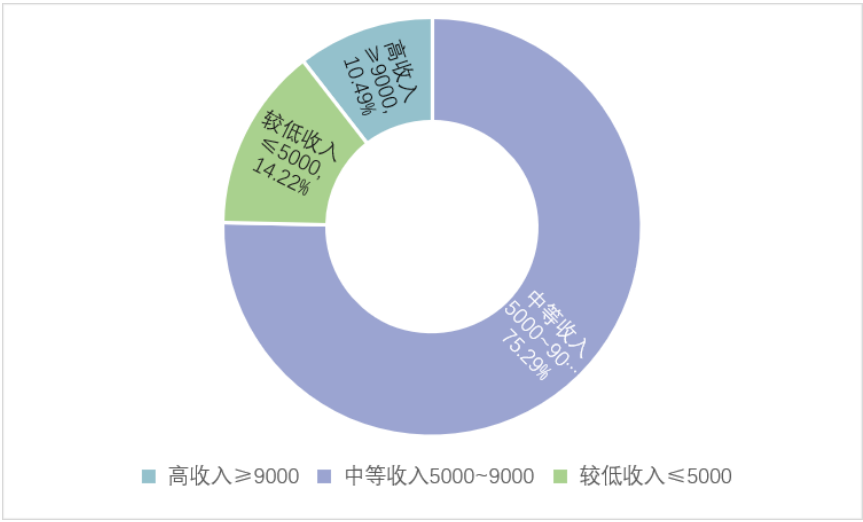


图 8　预期工资分类统计图

对家政从业者的预期工资，我们按标准进行划分，统计发现，预期工资处在中等收入水平的占主体，达 75.29%；预期工资处在较低收入的占比达到 14.22%；处在高收入占比达到 10.49%。从家政从业者对自己工资的制定可以看出，绝大多数制定的工资处在中等偏上水平，而工资更高对应的是更严格的人员要求，这些体现在学历、技能、经验等多方面。家政从业者通过对自己更高的要求，使自己更加受客户青睐，从而使期望工资也达到更高水平。

四、客户对家政从业者的需求——基于词频分析和词云图

为了更加了解家政求职人员的核心竞争力，本文通过爬虫爬取了有关各家政求职人员的培训老师评价和工作经历，共爬取有效信息 17435 条。

（一）词频分析

本文运用 Python 软件，对有关**培训老师评价和家政人员工作经历**的相应内容进行分词，在对无关和乱序词语清除后，使用词频统计分析筛选并排序**词频 50 和词频 200** 以上的关键词：

1. 培训老师评价词频排序

表 4　老师评价词频排序

排名	关键词	词频	排名	关键词	词频	排名	关键词	词频
1	认真负责	3759	12	非常	175	23	护理	76
2	难得	3758	13	勤快	144	24	客户	75
3	责任感	3746	14	不错	138	25	家务	73
4	专业培训	1960	15	性格	131	26	喜欢	69
5	爱心	1716	16	保洁	120	27	责任心	55
6	高端	1605	17	做饭	89	28	面食	53
7	参加	314	18	经验	88	29	整理	52
8	可靠	311	19	干净	88	30	人品	52
9	放心	310	20	孩子	87	31	烧饭	50
10	比较	310	21	温和	86			
11	工作	230	22	擅长	81			

对词频排名前 400 的关键词绘制词云图，并按各关键词按所占词频权重与频率控制大小与强调程度。

2.家政从业人员工作经历词频排序

表 5　工作经历词频排序

排名	关键词	词频	排名	关键词	词频	排名	关键词	词频
1	经验丰富	5886	16	干净	524	31	喜欢	304
2	利索	5271	17	学历	480	32	搭手	298
3	保洁	2906	18	清爽	467	33	培训	285
4	做饭	2457	19	勤快	453	34	家常菜	282
5	责任心	2293	20	全能	450	35	性格开朗	282
6	高端	1820	21	整理	421	36	育儿	275
7	护理	1513	22	添加	389	37	住家	274
8	母婴	1154	23	简单	382	38	接送	272
9	家庭	1147	24	温和	378	39	带过	246
10	早教	990	25	嫂子	352	40	洗烫	246
11	育婴师	906	26	照顾	348	41	双胞胎	239
12	家务	894	27	证书	347	42	新生儿	239
13	面食	654	28	姐妹	339	43	带到	227
14	烧饭	613	29	别墅	325	44	川菜	208
15	煲汤	554	30	老人	322	45	身高	200

根据**词频排名前 379** 的关键词绘制词云图，并按各关键词按所占词频权重与频率控制大小与强调程度。

（二）词云分析

1.培训老师评价词云图

图 9　老师评价词云图

2.家政从业者工作经历词云图

图 10　工作经历词云图

从词频分析和词云图可以看出：

（1）家政行业从业者的专业水平是客户需求的首要因素。“经验丰富”占比最高，“学历”、“工作”、“证书”“手脚麻利”等词也占有一定比例，说明对于家政服务人员来说，专业技能仍是必备的核心素养和聘请的重要前提，客户更有意图雇用受过正规化专业岗前培训的家政从业者[9]，良好丰富的经验与工作经历是从业者被雇佣的较大加分项。

（2）家政行业从业者的个人品质是客户选择时的重要考量。“责任心”、“温和”、“爱心”、“可靠”、“勤快”等词占比较高。家政从业者的工作场所靠近客户的生活空间，甚至也几乎成为客户家庭的一份子，良好的个人修养和积极的生活态度能够带给客户极大地舒适度和满意度，有利于家政从业工作者的上岗就业以及工作职业的稳定性。

（3）家政行业发展细分化、专业化为客户提供了广阔的选择空间。诸如

“高端”、“专业培训”等词占比较高，说明目前家政市场行业繁多，不再处于传统的洗衣做饭等日常家务的局限范围，家政服务除了有照看孩子、护理老人外，更加具有高端私人定制式需求的模式[10]。这需要家政从业者不断完善提升整体素质，建立以专业化为核心的质量标准。

根据词频、词云分析结果，结合背景，可以总结出受市场青睐的家政从业人员应具备的四方面素质：

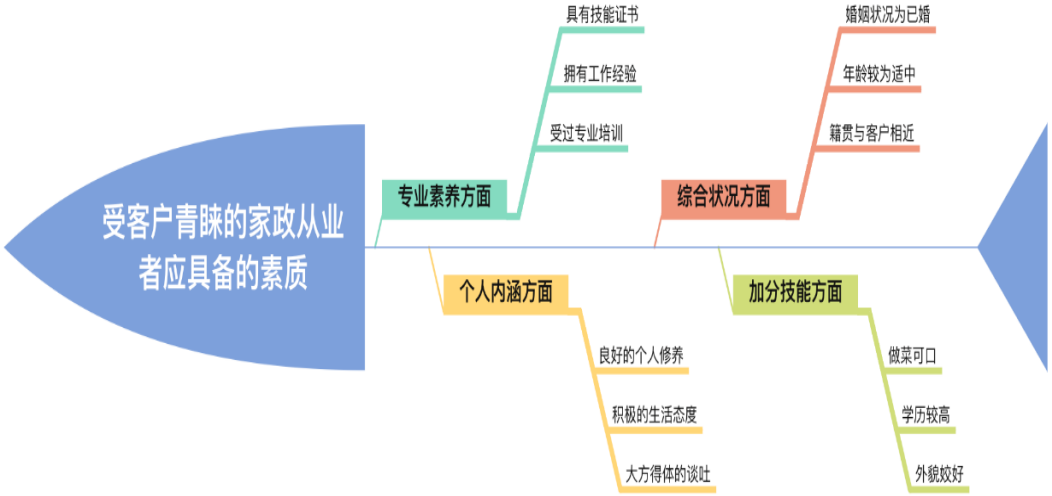


图 11　从业者具备素质鱼骨图

由图 11 可以看出，在家政服务需求持续上涨的同时，愿意为高品质家政服务买单的人进一步增多，客户对于家政服务人员的要求也不断提高，不仅要求专业技能的深度同时要求服务范围的广度，倒逼着家政服务行业的进一步发展。

五、家政从业者综合评估因子分析

（一）变量处理

我们对家政从业者的各项指标进行量化分析，处理标准与数据特征如表 6 所示。

表 6 家政从业人员综合评分影响因素的选择、定义与数据特征

变量名称	变量定义	均值	标准差	最小值	最大值
实名信息	实名=1，未实名=0	0.729	0.444	0	1
是否住家	住家=1，不住家=0	0.983	0.129	0	1
平均期望工资	个人期望工资范围的平均值	6721.322	2334.716	1000	80000
年龄	生存年数	46.103	6.361	20	86
学历	小学=6，初中=9，高中=12，中专=13，大专=15，大学=16，研究生=19	10.895	2.407	6	19
婚姻	已婚=1，未婚=0	0.994	0.076	0	1
籍贯	北京=1 ，天津=2 ，河北=3，山西=4，内蒙古=5，辽宁=6，吉林=7，黑龙江=8，上海=9，江苏=10，浙江=11，安徽=12 ，福建=13，江西=14，山东=15，河南=16，湖北=17，湖南=18，广东=19，广西=20，海南=21，重庆=22，四川=23，贵州=24，云南=25，西藏=26，陕西=27，甘肃=28，青海=29，宁夏=30，新疆=31，台湾=32，香港=33，澳门=34，国外=35	14.971	5.135	1	35
身材	BMI 值 18.5 以下=0，18.5-23.9=1，23.9 以上=2	1.073	0.322	0	2
经验	从事职业年限	6.811	17.759	0	38
做饭口味	会做菜系数量	4.444	1.953	1	18
会说语言	会说方言与外语数量	2.509	0.599	1	7
工作范围	工作内容数量	6.253	2.172	0	12

（二）因子分析

首先，通过计算 14 个变量之间的相关系数，即选取相邻两变量未缺损数据与各变量间进行计算，并对相关矩阵进行可视化以方便说明。

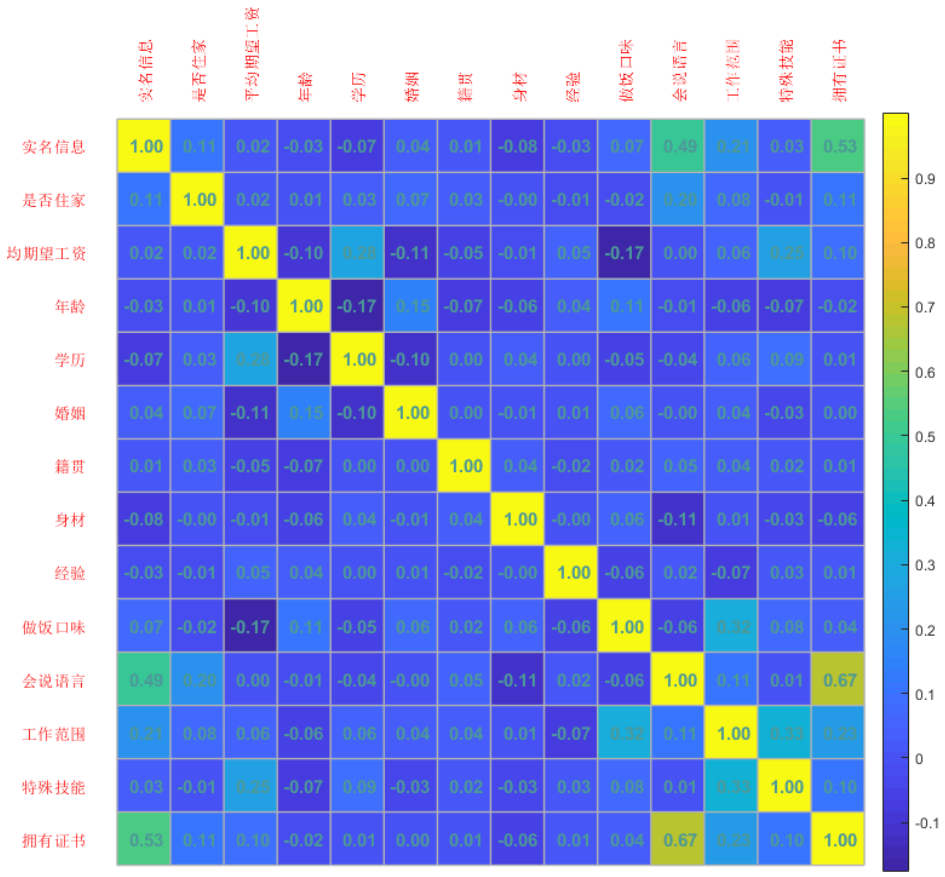


图 12　相关系数矩阵图

由图 12 可见，指标间的相关程度越强则颜色越浅，由于变量较多且有些变量间存在较强相关性。为使结果更加精确，我们考虑使用因子分析方法进行降维。

1.效果分析

接着我们进行了 KMO 和巴特利特检验，结果见表 7。

表 7　KMO 和巴特利特检验

KMO 取样适切性量数		0.636
巴特利特球形度检验	近似卡方	18546.714
	自由度	91
	显著性	.000

KMO 统计量为 0.636，变量间的相关性较强，偏相关性较弱，因子分析效果较好。=18524.714，显著性水平达到 0.000，数据呈球形分布，各个变量在一定程度上相互独立。

2.因子命名

在以上结论的基础上，我们进行了因子分析法。为使维度划分更加合理，我们通过主成分分析法，得出各项指标所占影响比重，具体结果如表 8 所示。我们提取出 5 个因子，并依据前文分析将这 5 个因子分别定义为专业水平、核心素养、服务能力、个人阅历与综合状况。

表 8　家政从业人员综合评分影响因素旋转正交因子表

指标	因子					命名
	1	2	3	4	5	
会说语言	0.402	-0.028	-0.117	-0.001	0.017	专业水平因子
拥有证书	0.378	0.024	0.017	-0.029	-0.050	
实名信息	0.343	-0.078	0.039	-0.055	-0.037	
平均期望工资	-0.002	0.503	-0.022	0.072	-0.123	核心素养因子
学历	-0.042	0.390	-0.023	-0.049	0.181	
工作范围	0.041	0.074	0.501	-0.005	0.045	服务能力因子
做饭口味	-0.068	-0.211	0.490	-0.009	0.044	
特殊技能	-0.045	0.319	0.364	0.057	-0.153	
婚姻	-0.027	-0.046	0.075	0.568	0.123	个人阅历因子
年龄	-0.042	-0.135	0.056	0.449	-0.264	
经验	-0.015	0.226	-0.129	0.416	-0.103	
身材	-0.090	0.022	0.031	0.038	0.532	综合状况因子
是否是流动人口	0.026	-0.060	0.000	-0.077	0.508	
是否住家	0.113	0.136	-0.093	0.407	0.436	

3.综合评分

根据旋转载荷平方和表，我们得到家政从业人员综合评分公式：

依据公式我们得出家政从业人员评分，分城市对评分与各项因子的平均值进行排序，我们得到如图 13 所示结果。

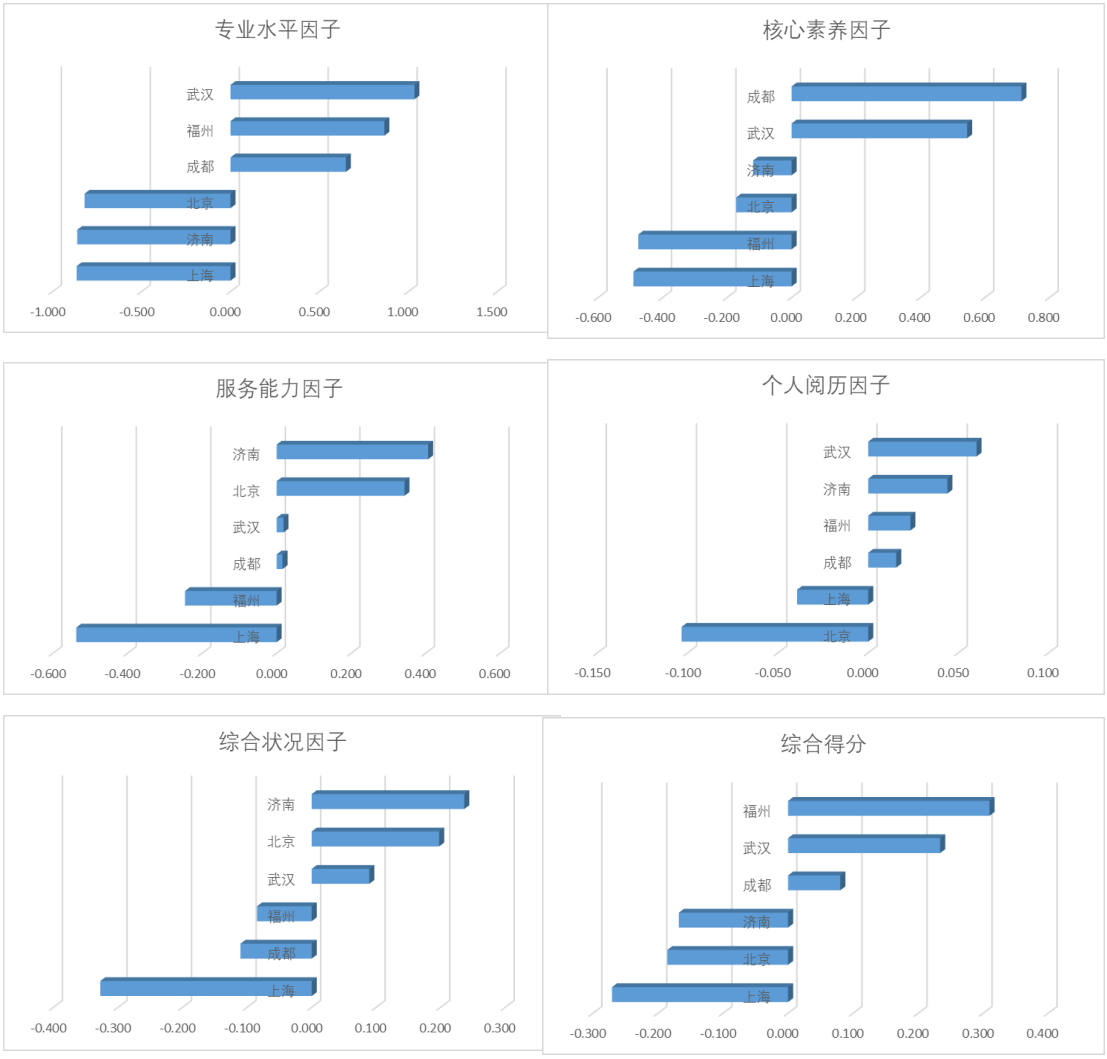


图 13　因子得分排名六宫格

我们由五个因子进行分析，对从业者进行评价时权重由高到低的因子排序为“专业水平”、“核心素养”、“服务能力”、“个人阅历”与“综合状况”。即影响从业者评价高低的重要指标为“会说语言”、“拥有信息”、“实名信息”、“做饭口味”、“学历”、“特殊技能”等。由此我们得出结论，从业者若是想要提高自己的总体水平与市场竞争力，首先需要考虑的便是提高自己的专业水平与服务客户的能力。

由城市排序我们进行分析，福州在综合得分中最为优秀，其家政从业人员

在专业水平方面的表现较为突出，在个人阅历方面表现较为良好，在核心素养、服务能力与综合状况方面表现较为一般，即福州家政行业更加规范，其从业者经验也更加丰富，但在专业水平上仍有待提高；武汉与成都在综合得分中较为优秀，主要在专业水平、核心素养与个人阅历方面表现较好，但在服务能力与综合状况上表现较为一般，即武汉与成都家政行业从业人员素质较高，但从业者个人资历方面较为欠缺；济南综合排名中等，在服务能力、综合状况方面表现较好，但其余表现一般；北京、上海两地在综合评分中表现较差，主要表现为专业水平、核心素养与服务能力水平较低，即北京上海家政从业者综合素质、总体水平与其他城市相比较低。

六、人员专业水平与预期工资的关系——回归分析

（一）变量选取

为进一步研究家政从业者的专业水平与核心素养对其制定预期工资的影响，在因子分析的基础上，我们选取了以下五个因子作为变量进行分析。

表 9　回归分析变量选取

指标名称	变量序号
专业水平因子	F1
核心素养因子	F2
服务能力因子	F3
个人阅历因子	F4
综合状况因子	F5

（二）回归分析

对回归分析模型进行解读，由表 10 可知，R 方的值为 0.609，即预期工资的 60.9%可以通过使用五个因子的回归分析来进行解读，模型拟合程度较好。

表 10　回归分析模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	0.780	0.609	0.609	1461.99007

对回归模型的方差结果进行分析，由表 11 可知，F 统计量为 3359.113，p 值为 0.000，小于 0.001，在 α =0.05 的检验下，拟合的回归方程具有统计学意义。

表 11　回归分析 ANOVA 方差分析表

	平方和	自由度	均方	F	显著性
回归	3.590E+10	5	7179817440	3359.113	.000
残差	2.306E+10	10790	2137414.962		
总计	5.896E+10	10795			

对回归系数进行分析如下表 12 所示。

表 12　回归分析系数表

变量	F1	F2	F3	F4	F5
系数	0.0152*** (0.00229)	0.207*** (0.0129)	0.000358 (0.00738)	-0.00403 (0.0197)	-0.0343*** (0.00621)

注：*表示在显著性为 10%，**表示显著性为 5%，***表示显著性为 1%，括号内表示标准误

依据表 12，得到回归方程：

$$\mathrm{Ln}\widehat{Y} = 8.771 + 0.0152 * F_1 + 0.207 * F_2 + 0.000358 * F_3 - 0.00403 * F_4 - 0.0343 * F_5$$

由该分析可以得出，预期工资的制定与家政从业者的专业水平联系密切，家政从业者所拥有的技能、服务质量与服务范围等都会在一定程度上影响她们的预期工资制定，且绝大部分变量与预期工资呈正比关系。若是家政从业者希望提高个人期望工资，他们的专业能力水平同样需要提高，这也会带动家政行业整体素质的提升与进步。

七、基于机器学习的预测——BP 神经网络

（一）问题分析

针对 “家政从业人员的预期工资是否能够达到行业的平均工资”的问题，根据网络上爬取的大量数据，我们优先选择进行机器学习，尝试建立 BP 神经网络模型并对其主要的 12 个影响因素进行分析。

目前人工神经网络智能算法中被最广泛应用的算法之一是具有反向传播功能的 BP 神经网络。该算法是由输入层、隐藏层与输出层三部分 [11] 组成的多层神经元，不同部分神经元具有不同作用，隐藏层数目未知，需设定。BP 神经网络模型不需要对数据整理与定量分析，具有较高的自我学习能力，具有通过自我训练来获得获取知识的能力，其预测精度较高、系统稳定可靠。

输入的信息经输入层、隐藏层后到达输出层，若输出结果与实际预期的结果有偏差，则将该偏差进行反向传输并通过修改权值和阈值使偏差变小，反复操作直到输出偏差满足预期要求，系统停止计算，并记录各神经元的权数，最终得到输出和预想结果一致的 BP 网络模型。

神经网络结构（三层隐藏层）示意图如图 14 所示。

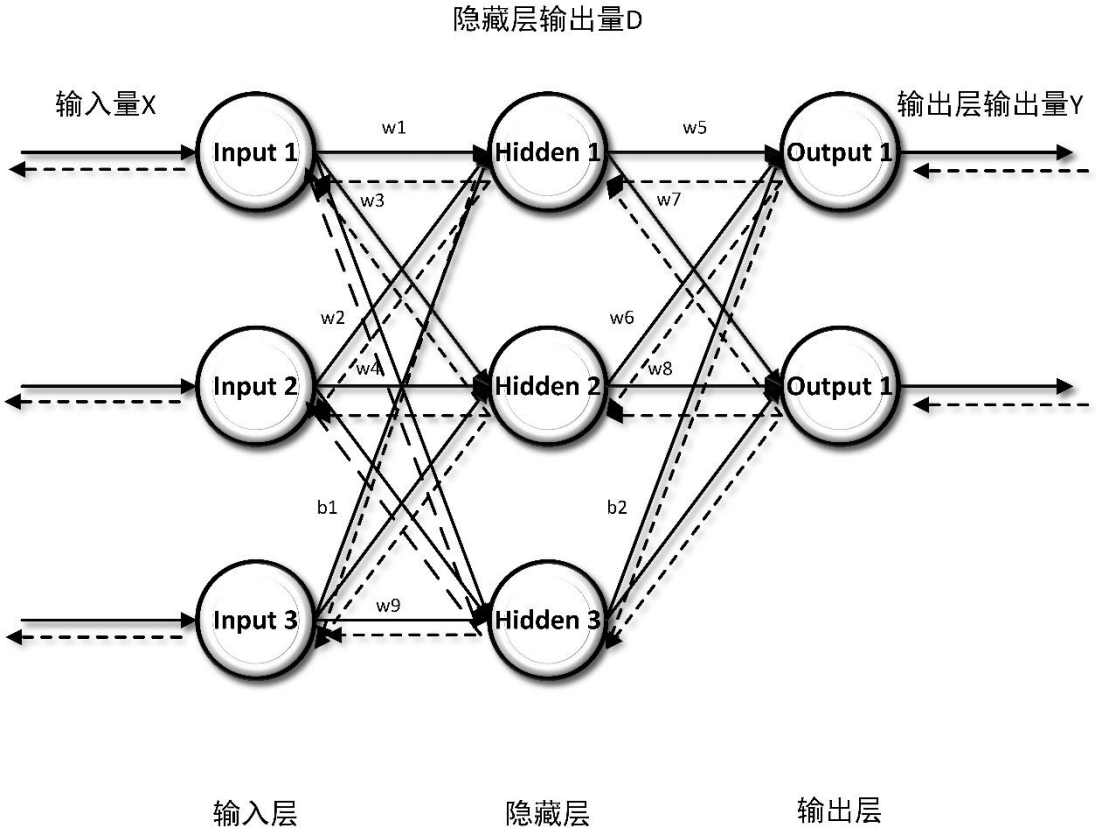


图 14　三层 BP 神经网络结构

（二）模型假设

1.假设爬取的数据真实可靠，且极端案例可忽略不计。

2.假设选取的六个城市具有很强的代表性与普遍性。

3.假设我们选取的隐藏层中神经元数目合适。

（三）对样本数据进行归一化处理

因为每一个数据的单位范围差异都很大，可能会产生训练时间较长与收敛速度较慢的情况。最终结果会导致波动范围大的数据作用会偏大，波动范围小，数据作用偏小。所以要使用归一化对样本数据进行处理。

采用 mapminmax( )函数对样本数据 X 进行处理，样本数据 Y 已经进行过 0-1 处理，就不做归一化处理。具体采用的公式如下：

$$x^{'} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

（四）模型建立

1.**在构建模型前先对数据进行预处理。**筛选与处理一共得到 10800 组数据。对工资水平进行 0-1 规划，将预期工资水平达到平均工资水平的设置为 1，未达到的设置为 0。将数据导入 MATLAB 的工作区。

2.**采用 MATLAB 内置的 Neural Net Fitting 进行仿真测试验证**。设置三层神经网络，并对数据分类、选择。共 13 个神经元，将是否能够达到预期工资产生影响的 12 个因素——实名信息、是否住家、年龄、学历、婚姻、籍贯、身材、经验、做饭口味、会说语言、工作范围、特殊技能、拥有证书分别定义为 X_1，X_2，.....X_{12}，并设置输入层为第一层；第二层设置为隐含层，其节点数待定；将“是否达到预期工资”的 0－1 变量 Y 设置为该神经网络的第三层，即输出层。

3.**划分训练集、确认集与测试集**。将预处理过的数据集随机选取 70%（7561 组）作为训练数据（Training）训练网络，10%（1080 组）作为确认数据（Validation）确认神经网络的训练效果，20%（2160 组）作为测试数据（Testing）判断神经网络的好坏。

4.**选择隐藏层中的神经元数目**。若隐藏层中的神经元数目量过多，会因其对其他数据识别能力较差而导致过拟合，反之则可能因为模式分类不够准确出现拟合效果不好 [12]。隐藏层节点数一般由公式（l 为隐藏层神经元个数，分别为输入输出层神经元个数，为 1~10 之间的数）确定，经过反复训练，本文采用 6 个隐藏层神经元，故 BP 神经网络结构为 12-6-1 [13]。

5.**样本训练**。采用 Levenberg-Marquardt 训练算法多次仿真测试，得到准确性较好的一次仿真测试，本次共进行了 35 次迭代。

（五）仿真结果与分析

图 15 是程序仿真时的界面图，表 13 是程序仿真结果。

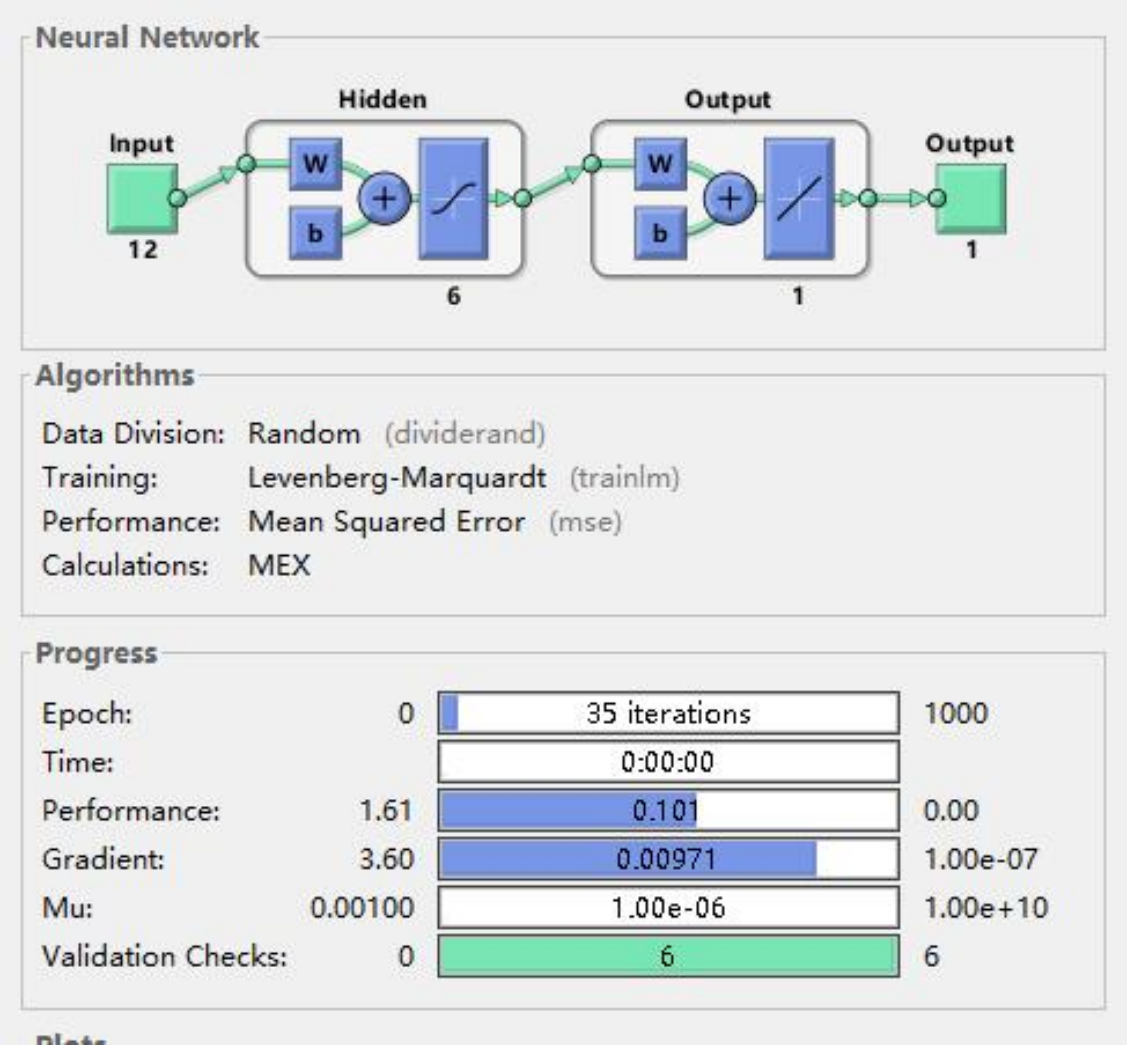


图 15　仿真程序界面图

表 13　仿真结果

	Samples	MSE	R
Training	7561	0.104906	0.614658
Validation	1080	0.993662	0.599122
Testing	2010	0.109751	0.577439

其中 MSE 是训练集、确认集与测试集的均方误差，R 是其对应拟合程度。图 16 是误差直方图。

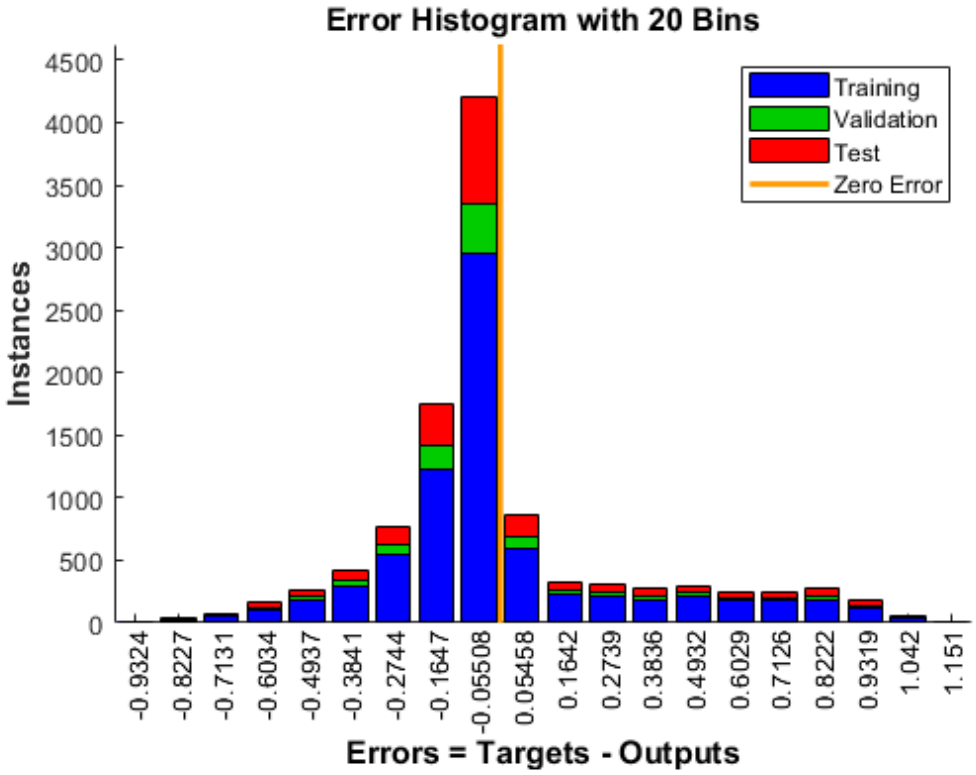
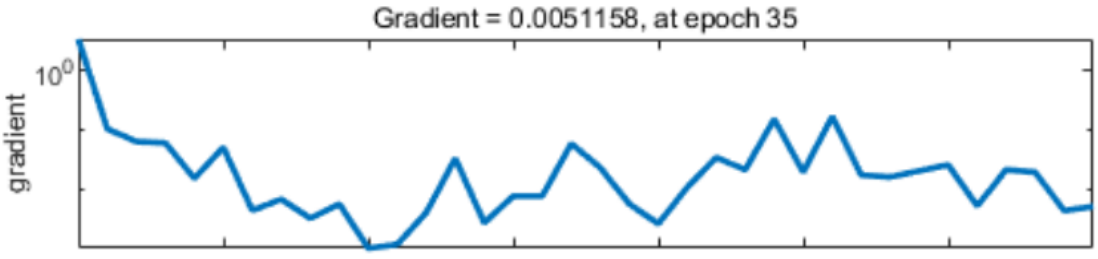


图 16　Error Histogram 图

由误差直方图可以看出，绝大部分误差在-0.05508 与 0.05458 之间，误差随着数据集的数目增大而减小，所有数据集的误差的范围都在-0.9324 与 1.15 之间。由于绝大部分误差在-0.05508 与 0.05458 之间，可见模型是比较准确的。



(a)

图 17　Training state 图

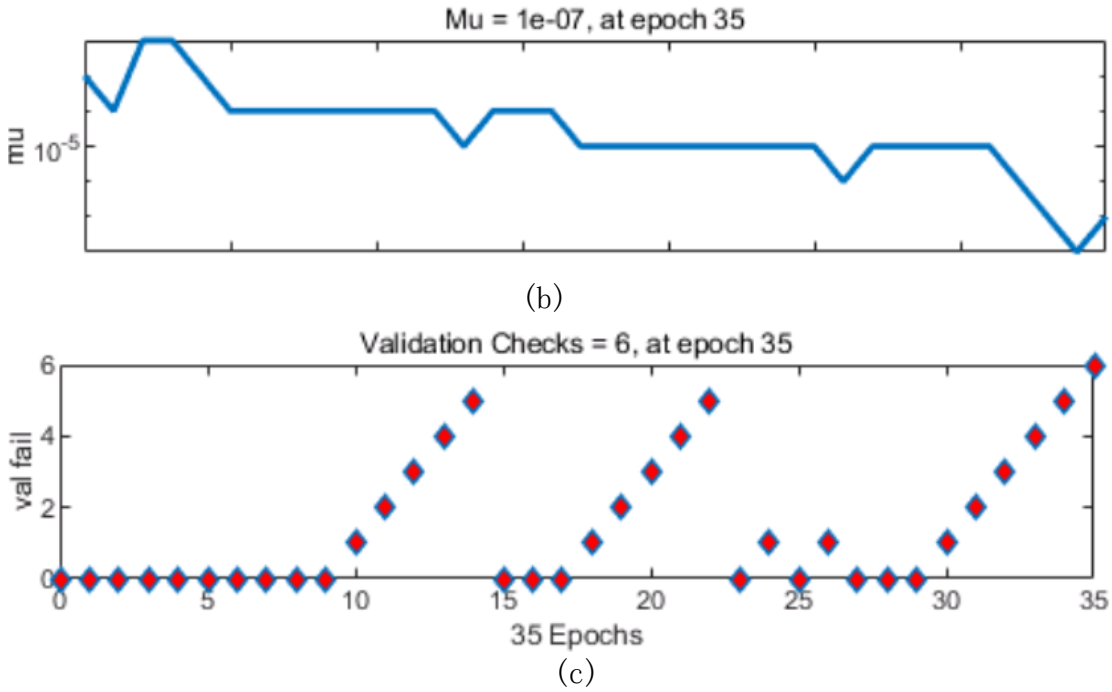


图 18　Training state 图

由 Training state 图可以看出训练状态，(a)图为梯度变化情况，可见 BP 神经网络算法的本质是按梯度变化的，目标函数的优化十分复杂，如图，梯度整体上是先下降再上升再下降的过程，具有“锯齿形现象”的特征。(b)图为 Mu 值的变化，整体成下降趋势。(c)图为校验检查图，仿真模拟预设经过六层检验结束运行，在第 35 次迭代结束。

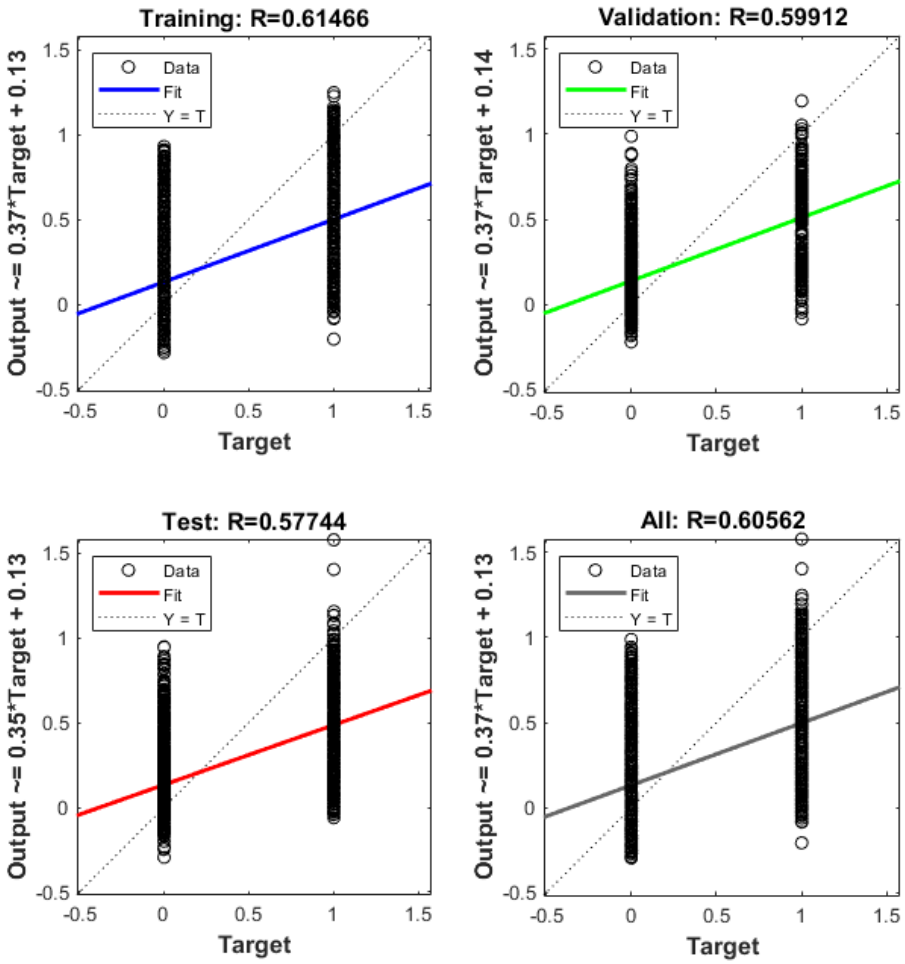


图 19　检验 R 指标值的 regression 图

该图检测了 R 值，反映了回归的拟合情况，每个 R 值都在 0.6 左右，且总体的 R 值为 0.60562。由于 R 值的范围在 0 到 1 之间，且 R 值越接近 1，预测值越准确，图中可见，训练集、验证集、测试集以及总体都是在目标为 1 时输出数据多数为 1，在目标为 0 时输出数据多数为 0，而且训练集与测试集分布状况相近，因而可以判断出该神经网络模型性能较为良好。

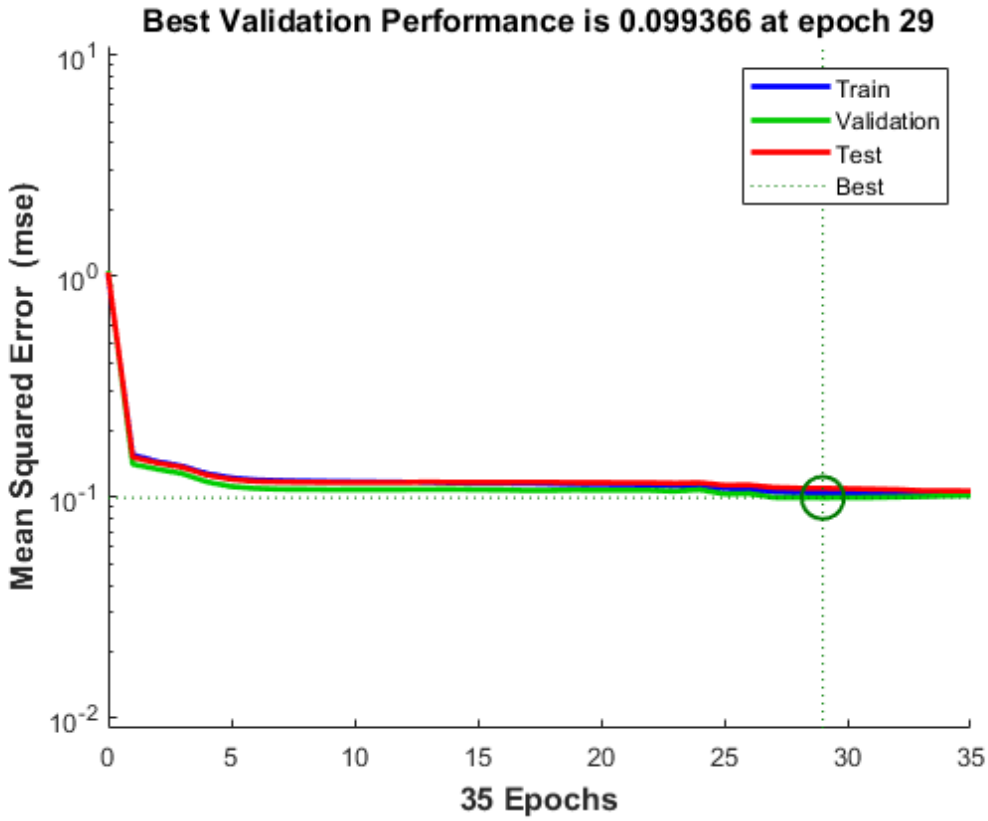


图 20 Performance 图

由 Performance 图可以看出，神经网络在第 29 次迭代达到最佳性能，停止训练且训练网络 Validation Checks 指标达到设置值 6，网络误差为 0.099366 ，即准确率为 90%左右，达到了对网络模型性能准确性的预期要求。可见，训练集与测试集变化趋势一致，训练次数的增加，训练集与测试集误差逐步减小。当网络训练到 35 次后，3 条曲线基本变成一条，误差逐渐稳定 [14]，可以看出训练集与验证集几乎重合。

通过前面的研究得知 BP 神经网络模型具有一定的准确率，但是该算法的收敛速度慢、局部搜索优化等问题导致其不能达到预期的满意度，因而决定使用模型筛选器 Classification Learner 仿真模拟，以期选出合适的模型。

八、根据因素对预期工资是否达到平均工资的判断——决策树模型

（一）模型的筛选

为了使机器学习的模型的选择达到最优，我们使用 MATLAB 内置的模型筛选器 Classification Learner 仿真模拟。在导入爬取的 10800 组数据组成的数值矩阵后，将预期工资设置为响应器（responsor），其余 12 个元素为影响因素，设置为预报器（predictors），本文根据样本数量选择 10 折交叉检验法分别交叉验证。在选取筛选器中所有的模型后，开始仿真模拟。

选取性能最佳的支持向量机、k-近邻算法、逻辑回归法、袋装树、提升树、优良树种六种模型。六组模型的精确度与 AUC 值如下表。通常选择综合性较高的 AUC 值作为模型好坏的判别标准。AUC 值为 ROC 曲线下的面积，其值越接近 1 表明模型的精度越高。

表 14　模型精确度表

	精确度（Accuracy）	接受者操作特征曲线 AUC 值（Area under the Curve）
介质高斯支持向量机（Medium　Gaussian　SVM）	86.5%	0.84
余弦 k-近邻算法（Weighted　KNN）	86.8%	0.88
逻辑回归法（Logisitic　Regression）	83.3%	0.84
袋装树（Bagged　Trees）	87.6%	0.90
提升树（Boosted　Trees）	86.5%	0.88
优良树种（Fine　Trees）	85.9%	0.85

由表 14 可以看出， 6 个模型中最适用的是袋装树模型，其精确度达到了 87%，AUC 值达到了 0.90，图 21 为其混淆矩阵图，在图中第一行，96%的预测工资未达到平均工资的数据被正确分类，在第二行中，有 57%的预测工资达

到平均工资的数据被正确分类，即有 43%的数据被分类到了未达到平均工资的数据。可见该决策树模型对预测工资未达到平均工资的数据的判断具有很高的正确率。

图 21　混淆矩阵（Confusion Matrix）

综上，袋装决策树模型具有很高的准确性。本文建立袋装决策树模型进行分析，其模型的生成及决策共有三个模块：（1）训练集经过递归分析后生成倒立的树；（2）分析从树的根节点开始到叶子节点结束的路径，以形成规则；（3）根据这些规则对新数据进行分析与预测。

（二）构造决策树模型

1.利用 MATLAB 编写决策树程序，将归一化处理的数据导入后运行，按照 80%选取训练集 8640 组，测试集为 2160 组。对训练集进行训练，得到决策树如图 22 所示，可见未优化、未剪枝的决策树模型结构复杂，且运行时间长，因此，要对其改进。

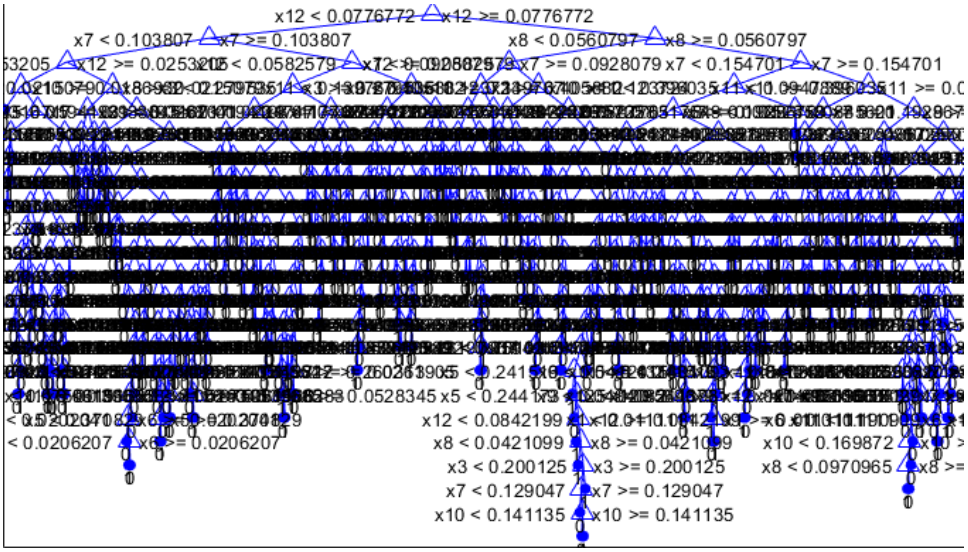


图 22　未优化、未剪枝的决策树模型图

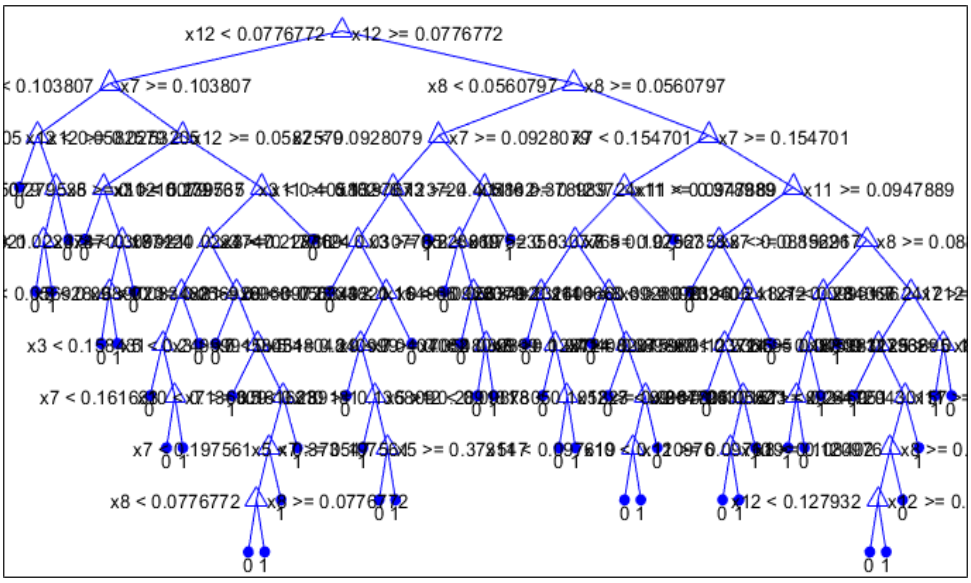


图 23　优化后的决策树模型图

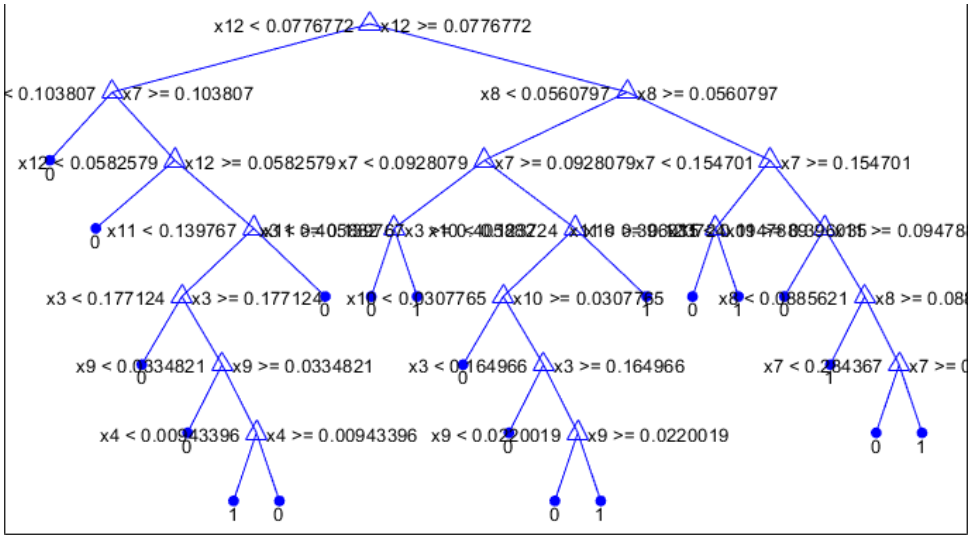


图 24　剪枝后的决策树模型图

2.对测试集进行仿真模拟，从表 15、表 16 可以看出该模型对未达到预期工资判断的正确率为 0.8882，对达到预期工资判断的正确率为 0.61522。

表 15　运算结果表 1

	未达到预期工资	达到预期工资	样本总数
样本集	8525	2275	10800
训练集	6825	1815	8640
测试集	1700	460	2160

表 16　预算结果表 2

未达到预期工资人数	误判数	正确率 p
1511	189	0.8882
达到预期工资人数	误判数	正确率 p1
283	1417	0.61522

（三）优化决策树模型与剪枝

1.**优化决策树**。清楚叶子节点所包含的最小样本数对其性能的影响后可优化决策树，本文采用交叉验证法验证误差 [15]。

通过实验知当叶子节点的最小样本数为 60 时，交叉验证所得误差最小，误差为 0.1478。将叶子节点的最小样本数设为 60。优化后的决策树如图 23 所示，其枝叶明显减少，拟合程度降低。其重采样误差与交叉验证误差从优化前的 0.0507、0.1747 变为了 0.1215 和 0.1515 [16]。

2.**对决策树剪枝。**对决策树剪枝，得图24的决策树。剪枝后的决策树为最终的决策树，其结构较为简洁，程序运行时间大大缩短，能达到预期设置。但其重采样误差为0.1334，交叉验证误差为0.1735，相较剪枝前略提高，拟合度相较剪枝前略降低。

九、影响因素对预期工资是否达到平均工资的重要性分析——随机森林模型

（一）问题分析

通过前面的研究分析得知剪枝后的最终决策树，结构较为简洁，程序运行时间大大缩短，能达到预期设置，但其重采样误差及交叉验证误差，相较剪枝前略提高，拟合度相较剪枝前略降低。因而本节在决策树模型的基础上，使用 MATLAB 的内置函数 TreeBagger（）依靠 Classification Tree 和 Regression Tree 功能来生长单棵树，再用多颗决策树建立随机森林（Random Forest，RF）模型，根据少数服从多数的投票原则对“家政从业人员的预期工资是否能够达到行业的平均工资”做判断，并分析其各个影响因素的重要性程度。

随机森林是一种具有监督作用的集成学习算法，以决策树为基本模型的袋装分类法模型 [17]。由于其抗干扰能力好、不易过拟合等优点，所以比决策树模型应用更广泛。一般由以下三部分组成：

1.**根据 bootstrap 自主采样法采样**。对 n 个样本的原始数据进行 t 次随机抽出有放回的采样，得到含有 t 个样本的采样集。如此反复操作，最终获得 N 个含 t 个训练样本的采样集 [18]。

2.**建立基决策树。**每个采样集建立一棵基决策树，每个节点的属性是先从该节点的属性集合中随机的选取一个包含 k 个属性的子集，再从这个子集中选择最优属性。

3.**分类准则。**建立 N 棵决策树后，遵循少数服从多数的投票原则，得出随机森林的分类结果。

决策树在被训练后会对“是否达到预期工资”这一问题具有一定的判断力，所有决策树会根据自己被训练后所具有的判断能力对问题进行投票判断。由于

随机森林在节点分裂时会随机选择任意某几个属性参与比较，每个决策树的投票准则都是不同的，采取大多数投票法实现算法最终结果输出[19]，在分类汇总后，被投票数最多的结果会作为算法的最终输出结果[20]。流程如图 25 所示。

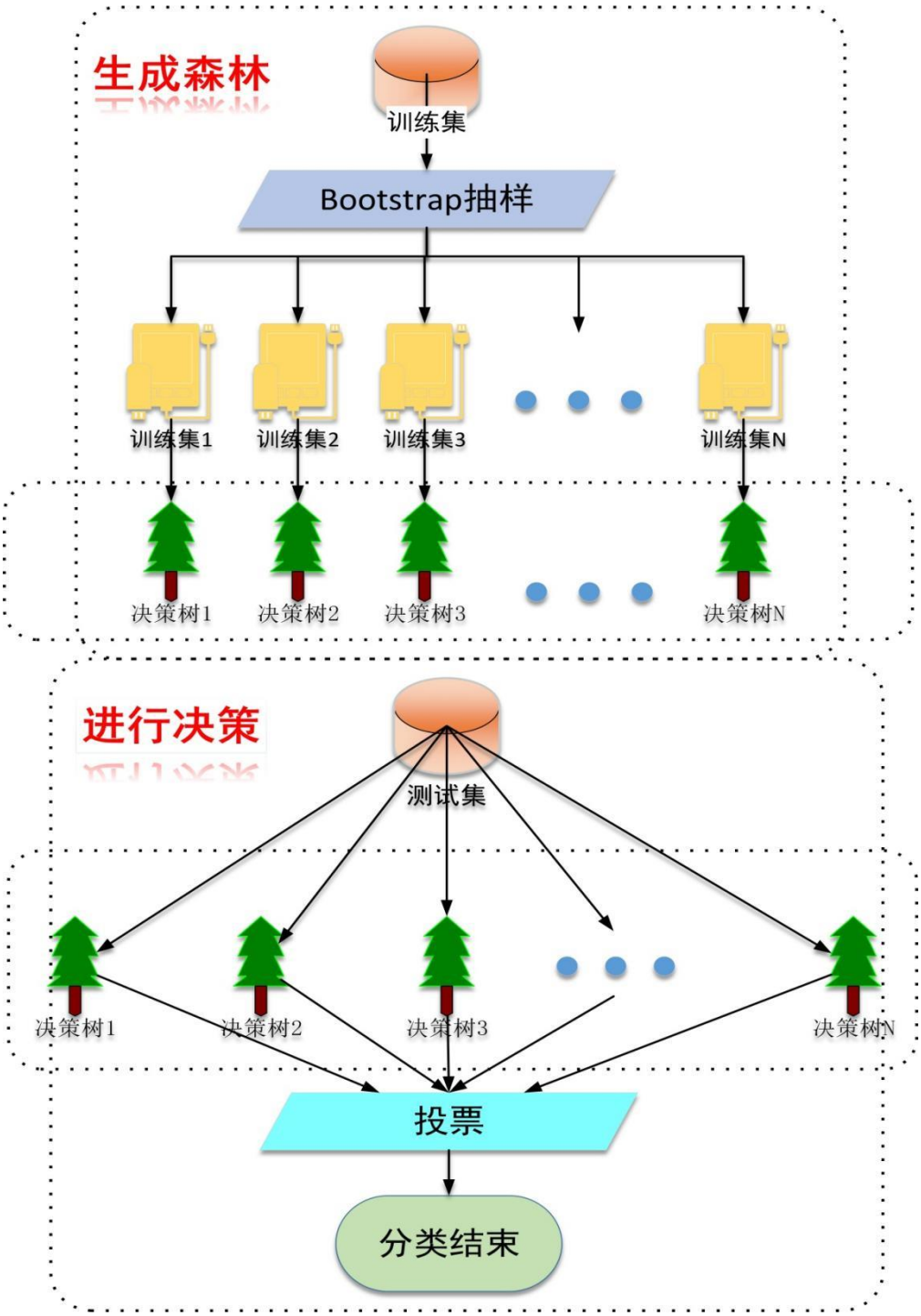


图 25　随机森林（Random forest）算法示意图

（二）模型假设

1.同前文 BP 神经网络模型假设 1、2。

2.假设随机森林模型选取的树都不做剪枝处理。

3.leaf 值为 23 时性能最佳。

（三）样本归一化处理

采用 mapminmax( )函数，方法同 BP 神经网络模型中一致。

（四）模型建立

1.删除无效和缺失的数据，一共得到 10800 组数据，为 108003 的矩阵。将数据导入 MATLAB 的工作区，定义为数值矩阵 F。

2.采用 MATLAB 的内置函数 TreeBagger（）建立随机森林模型。设置 leaf=23，决策树数目 500，函数编写完成后将 F 输入并运行。

（五）模型运行结果与分析

程序运行后，根据结果分别输出模型相关系数图、输入变量重要性图与 OOB 图。现根据这三张图分析输出结果与模型的优良程度。

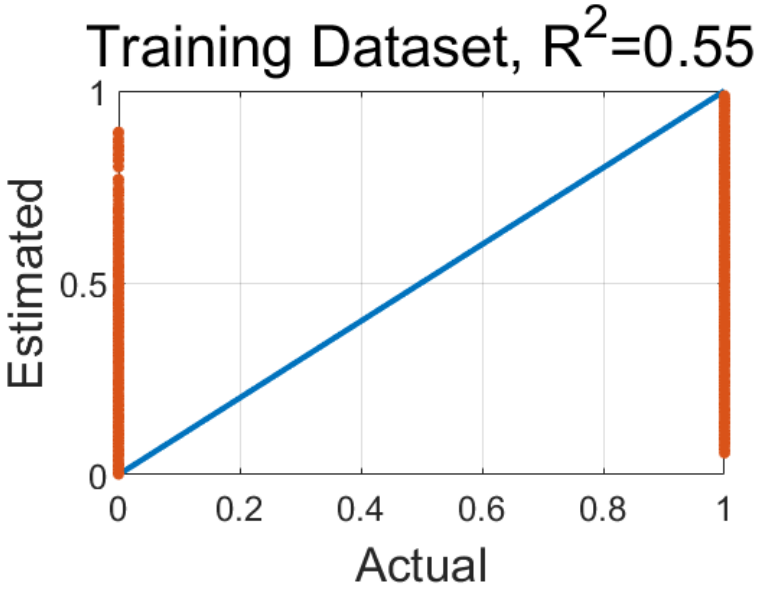


图 26　模型相关系数图

图 26 是训练集的真实值与估计值的分布图，真实值成 0-1 两点分布，估计

值在 0 和 1 之间分布，且对真实值为 0 的估计值虽有误判但未判断为 1，对真实值为 1 的估计值虽有误判但未判断为 0。该模型的 $R^2 = 0.55$，较为准确。

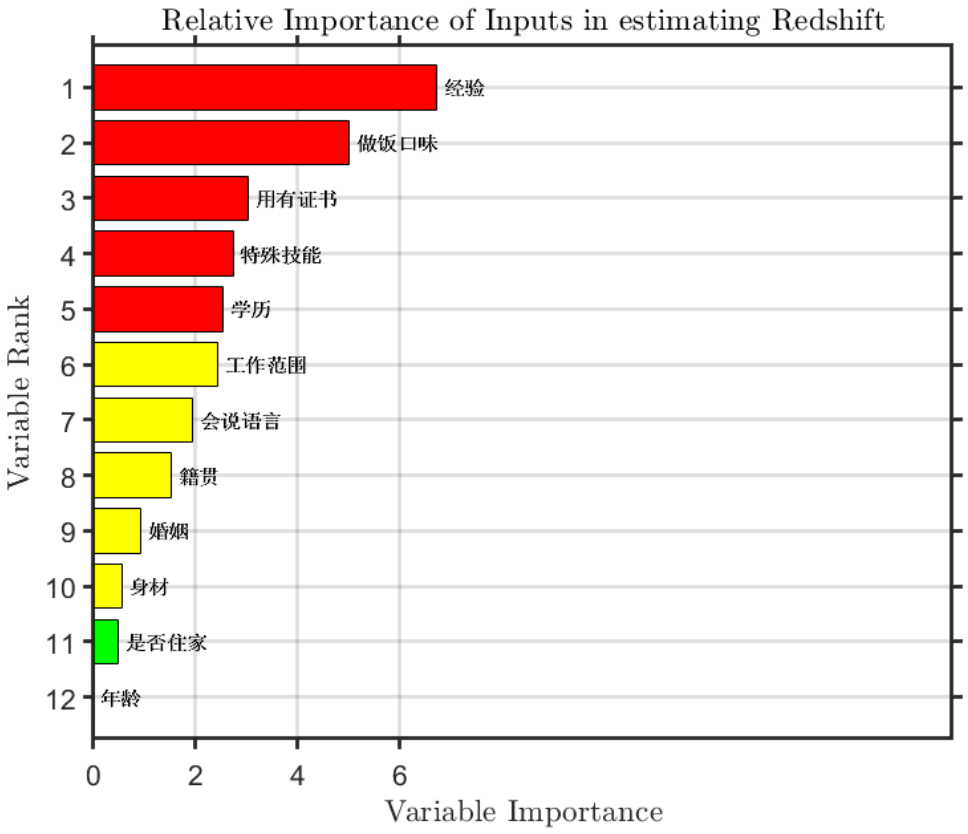


图 27　输入变量重要性图

输入变量的重要性程度示意图如图 27，可见经验与做饭口味是家政行业从业者“预期工资是否能够达到平均工资”的最重要的两个影响因素；由图可见年龄的重要性为 0，得出年龄可以不作为该问题的影响因素。重要性程度前三的因素分别来自个人阅历因子、服务能力因子和专业水平因子。

重要性程度具体值见表 17，其数值的大小代表重要性程度的大小，数值越大影响因素的重要程度越大，反之亦然。为了更为直观的了解各因素影响力大小的关系，我们利用 Tableau 将数据可视化，建立了气泡图，数值各因素影响力气泡图见图 28。，各因素的影响力大小与气泡的大小、色彩亮度成正比。

由上述三张图表，我们能够分析出家政从业者如何快速提升自身综合素质的“捷径”，能够为家政从业者提高自身在行业中的竞争力提供对策。

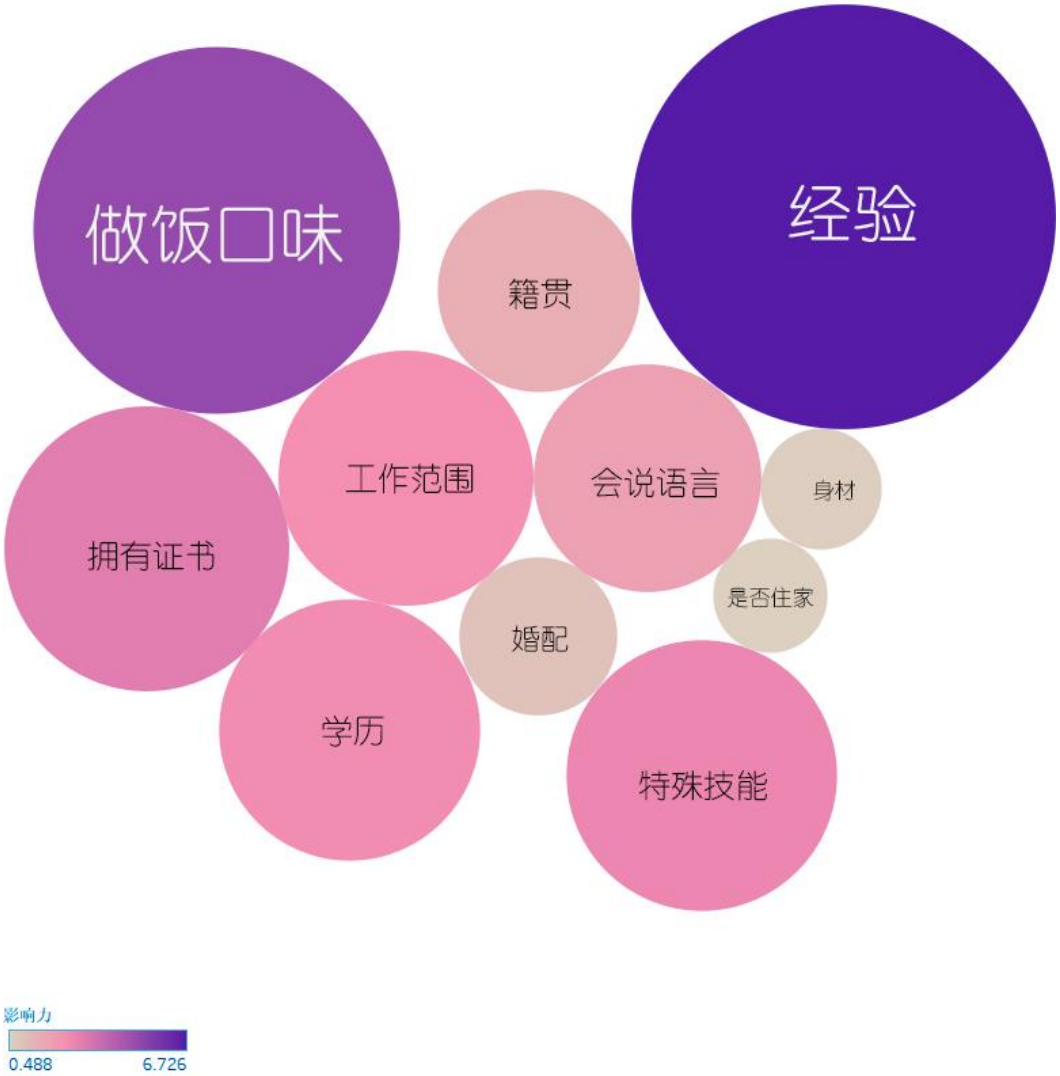


图 28　各因素影响力气泡图

表 17　各因子重要程度

影响因素	重要性程度	重要性排名	影响因素	重要性程度	重要性排名
经验	6.726	1	会说语言	1.934	7
做饭口味	5.015	2	籍贯	1.532	8
拥有证书	3.029	3	婚姻	0.9337	9
特殊技能	2.733	4	身材	0.5478	10
学历	2.545	5	是否住家	0.488	11
工作范围	2.429	6	年龄	0	12

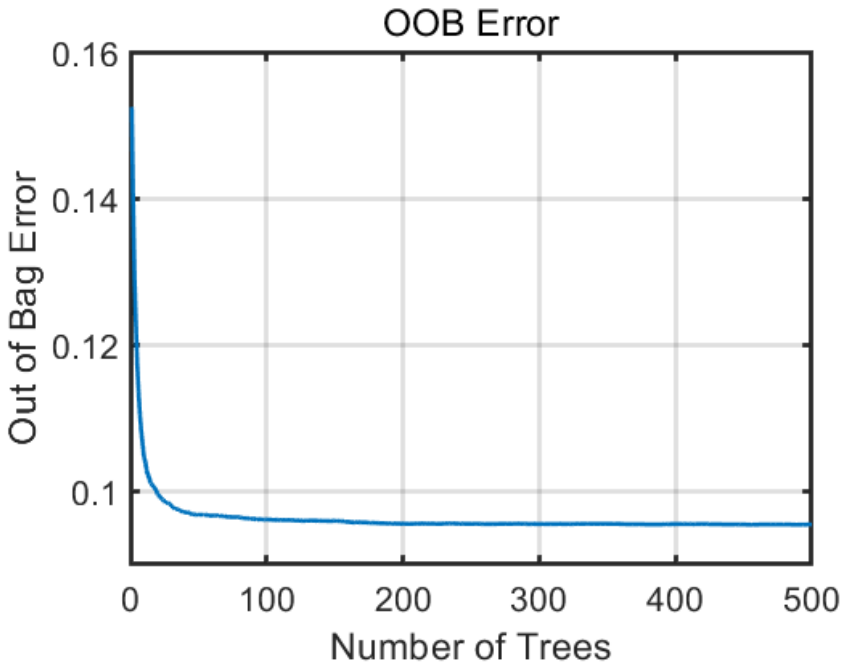


图 29　决策树数目（Number of Trees）与袋外错误（OOB Error）

随机森林在建立的时候，有一部分样本不在所采集的样本集合中，并未参与随机森林的建立，这些袋外数据 OOB（Out Of Bag）可代替测试集误差估计方法。用袋外数据测试已生成的随即森林其性能：假设 OOB 总数为 M，输入带进前已经生成的随机森林分类器，得出 M 条相应的分类。已知 M 条数据的类型，用正确的分类与随机森林分类器的结果做比较，记录筛选错误的数目，将其计为 N，则 OOB 误差大小为。由于其已知为无偏估计的，则不用再通过验证或建立测试集获取测试集误差的无偏估计。我们设置随机森林中的有 500 棵决策树，随着决策树的增加，袋外错误率越来越低，最后趋于平稳。 OOB Error 趋于 0.09543，模型准确性约为 90.5%。

随机森林模型由多颗袋装树组成的，其独特的生长分类方式与投票机制，使得其在本文的研究目标上具有非常高的适配性。

十、结论与建议

（一）结论

1.回归分析

本文主要研究了在中国家政市场规模不断扩张的背景下，为应对行业竞争压力与人工智能家居产业发展的冲击，家政行业从业者提升个人核心素养与专业服务的综合能力以加强个人市场竞争力[21]，并达到实际工资达到预期工资的目的，进而带动家政产业整体的进步。主要结论如下：

（1）家政行业从业者主要来自中部省份，以中年女性为主体，学历主要以初中学历为主，占绝大多数的求职意向对技能要求较低。

（2）客户选择家政员标准以专业水平为主，以个人水平为辅；家政行业发展细分化、专业化为客户提供了广阔的选择空间。

（3）影响每个家政从业者的综合评价更高的主要因素有五方面：①**专业水平**。家政从业者专业的高低是衡量其水平的重要影响因素。②**核心素养**。家政从业者的学历水平高低等影响其核心素质，这会影响到客户的选择。③**服务能力的提高**。家政从业者提高其综合服务能力，会使在客户选择时优先考虑。④**个人阅历提升**。个人经验与人生阅历的丰富可以提升从业者市场竞争力。⑤**综合状况的影响**。从业者的个人状况如籍贯身材等也会影响综合得分的高低。

2.行业现象分析

根据随机森林模型结果、文本及数据分析，发现使家政从业者预期工资能够达到平均工资的重要因素反映了如下行业现象：

（1）家政从业者现状分析

①“行业老人”独占鳌头，依然发光发热。越卷越吃香，正如干中学，一个人一件事做得长了自然会有一定的水平，家政行业亦如此，经验丰富、做事

老道的“行业老人”大都有极高的工资预期，却仍炙手可热，这是很多拥有特殊技能的“行业新人”难以企及的。

②多才多艺，方可多财多亿。家政行业内卷严重，在其中发光发亮需有一技之长，有甚者竟一人独掌八国语言。掌握更多技能的人对工资的期望普遍高，但他们却更容易被雇佣，因为市场需求正是如此。

③英雄不问出处，能力大于外在，高收入有能力者得之。家政不是什么都卷，不看出身，不论美丑，只要业务能力出色，都能预期一份不错的收入。雇主花钱雇佣是为的是办事，不会看上“花瓶”。

④多一个证件，多一份收入。证书是必不可少的敲门砖，也是是否在卷的良好证明，丰富的证书使从业者自身价值提高，也是谁不喜欢更优秀的人呢？

（2）城市综合状况分析

北京、上海等发达城市虽然家政市场范围广、规模大，但层次差距也较为明显，人员素质良莠不齐，平均水平有待提高。武汉、福州等城市家政行业总体水平更高，家政行业高质量发展趋势较为明显。

3.模型分析

根据 BP 神经网络模型、决策树模型与随机森林模型结果进行分析，运用所给 12 个因素判断“家政从业者预期工资能够达到平均工资”的模型正确率分别为 90.000634%、86.66%、90.457%，又因为 BP 神经网络模型算法存在一定的收敛速度慢、局部搜索优化问题，所以**最优模型为随机森林模型**。

（二）建议

1.家政从业者角度

参与正规专业培训，丰富个人专业技能，提升市场竞争力。现阶段，家政服务正实现从劳务型向知识型转型，故家政从业者需提高个人职业素养与核心竞争力，以满足客户更高标准高质量要求。

2.家政企业角度

优化人力资源，提升服务质量。家政企业为了使客户享受更高水平服务，可以通过客户的需求来对从业者进行细致的岗前培训，提升家政从业者的核心竞争力。

3.政府角度

政府对家政服务的培训机构也应出台财政补贴等优惠政策，对家政服务企业发展、家政服务培训及家政服务业标准制定和网络建设等给予资金支持，使得培训机构进入市场。

（三）不足与展望

针对北京、上海等服务业发展水平较高却综合评分较低等状况，通过查阅相关家政从业者的访谈资料，我们发现跳槽率高和拥有相关资历的家政人员少是目前家政行业发展的重要阻碍。且上海、北京等城市家政的需求缺口大和工资高（相对其他城市）更能吸引很多家政从业者跳槽前往。长此以往，在微观角度上降低当地家政从业者的工作压力，使其家政从业环境更为轻松，在宏观角度上很可能会阻碍北京、上海等大城市的家政行业高质量发展。

事实上，全国家政消费确实存在明显的地域差异，如全国消费者中使用家政保洁服务最为频繁的城市为北京；全国最爱搬家的城市是上海。于是，为进一步挖掘各城市家政行业发展排名背后的原因，我们团队还尝试搜集了《中国家政服务市场现状调研与发展前景分析报告（2021-2027 年）》、《中国家政服务市场需求调查及未来发展规划分析报告 2016-2021 年》等研究咨询报告，这些报告涉及武汉、济南、北京等重点城市的家政服务行业市场调研，但因这些研究报告价格昂贵和比赛时间原因而暂时止步，今后我们团队会进一步深入挖掘各城市背后的家政故事。

参考文献

[1]李艳梅.我国家政服务业的现状分析与规范化建设[J].社会科学家,2008(07):107-110+113.

[2]曹华.对我国家政服务业发展的一些思考[J].经济师,2003(04):61-62.

[3]苏明,梁季,唐海秀.我国发展家庭服务业促进就业的财税政策研究[J].经济研究参考,2010(52):2-14.

[4]孙学致,王丽颖.我国家政服务业规范化发展问题研究[J].经济纵横,2020(05):115-120.

[5]谷素萍.家政服务标准化建设和质量提升路径研究[J].人民论坛,2019(27):80-81.

[6]杨军剑.我国家政服务质量存在的主要问题及对策建议[J].经济研究导刊,2021(04):147-150.

[7]张贝妹,冯玉珠.后疫情时代河北省家政服务业发展的路径和对策研究[J].经营与管理,2021(06):184-187.

[8]尹航,夏凡.中国家政服务行业的现状、问题及对策探索——基于河南省家政服务公司发展经验思考[J].商业文化(上半月),2011(09):196-197.

[9]李银雪.上海家政服务业发展调研报告[J].科学发展,2021(05):108-113.

[10]李艳梅.我国家政服务业的现状分析与规范化建设[J].社会科学家,2008(07):107-110+113.

[11]王勃,崔洋,董丽欣.基于随机森林算法的高层建筑机械拆除方法判断[J].低温建筑技术,2020,42(12):40-42+46.

[12] 莫智焱. 基于机器学习的房贷决策引擎的设计与实现[D].哈尔滨工业大学,2018.

[13]王战,田婧,FENG Chuan.“工具理性”与“价值理性”的博弈:关于“雇主-家政工”关系的一个分析框架[J].华中科技大学学报(社会科学

版),2020,34(05):133-140.

[14]孙学致,王丽颖.我国家政服务业规范化发展问题研究[J].经济纵横,2020(05):115-120.

[15]史钰斌. 基于多策略的高校助学金精准预测的研究与应用[D].南昌大学,2019.

[16] Tahani Daghistani,Riyad Alshammari. Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes - Volume 11, No. 2, May 2020 - JAIT[J]. JAIT,2020,11(2).

[17]Horst Heather,Sinanan Jolynna,Hjorth Larissa,Horst Heather,Sinanan Jolynna,Abed Y,Chavan M,Akiti L,Andrejevic M,Hearn A,Kennedy H,Appadurai A,Beer D,Burrows R,Bell G,Dourish P,Buchli V,Clark LS,Goggin G,Hjorth L,Dalton CM,Taylor L,Thatcher J,Davison HK,Maraist C,Bing MN,Dion D,Sabri O,Guillard V,Erstad O,Sefton Green J,Furlong K,Gregg M,Grinter RE,Edwards WK,Newman M,Haldrup M,Larsen J,Hargittai E,Marwick A,Horst H,Horst H,Miller D,Horst HA,Taylor EB,Hsin Hsuan MN,Gneezy A,Griskevicius V,Williams P,Hu T H,Iliadis A,Russo F,Ito M,Okabe D,Anderson K,Ling R,Campbell S,Jackson SJ,Edwards PN,Bowker GC,Keane W,Myers F,Keane W,Kennedy J,Nansen B,Arnold M,Kirk DS,Sellen A,Lobato R,Thomas J,Löfgren O,Longhurst R,Lupton D,Lupton D,Lupton D,Lyman P,Varian HR,Lyman P,Varian HR,Lyon D,Marcoux JS,Marcoux JS,Miller D,Marwick A,Mehta R,Belk RW,Myers F,Myers F,Nippert Eng CE,Nippert Eng CE,Pickren G,Pink S,Leder Mackley K,Pink S,Lanzeni D,Horst H,Pink S,Sinanan J,Hjorth L,Pink S,Sumartojo S,Lupton D,Sertã AL,El Mimouni H,Barkhuus L,Shove E,Silverstone R,Hirsch E,Morley D,Hirsch E,Silverstone R,Sterne J,Sterne J,Strengers Y,Maller C,Strengers Y,Taylor AS,Harper R,Tolmie P,Crabtree A,Rodden T,Bannon L,Wagner I,Gutwin C,Urry J,Larsen J,Van House NA,Davis M,Ames M,Villi M.

Digital housekeeping: Living with data[J]. New Media & Society,2021,23(4).

[18]潘梦婷. 引进红利因子和企业质量因素的智能选股策略研究[D].上海师范大学,2019.

[19]纪思琪,吴芳,李乃祥.基于决策树的蔬菜病害静态预警模型[J].天津农学院学报,2017,24(02):77-80.

[20]彭广盼. CVD 金刚石涂层钻头加工钼圆孔技术的研究[D].江南大学,2013.

[21]黄驿惠. 基于改进 BP 神经网络的公交到站预报修正方法研究[A]. 中国智能交通协会.第十四届中国智能交通年会论文集（2）[C].中国智能交通协会:中国智能交通协会,2019:6.

致谢

在论文完成之际，我们要特别感谢指导老师。老师在选题、文献搜集、模型构建与编写等方面都耐心帮忙，悉心教导，给予了我们很大的支持与帮助！在此，我们向老师表示真挚的谢意！有幸得到她的指导，我们的论文才能够顺利撰写完成。

同时，也感谢小组成员的努力与坚持，是大家共同的付出，才使得论文撰写圆满完成。

最后，向在百忙之中抽出时间对本文进行评审并提出宝贵意见的专家表示衷心的感谢！