大连理工大学软件学院科技写作作业

基于多元线性回归模型的
软件行业薪资影响因素分析

An analysis of factors influencing salary in software industry based on multiple linear regression model

学 院（系）：软件学院

专　　业：软件工程

学 生 姓 名：张一骏

学　　号：201892272

完 成 日 期：2020 年 7 月 6 日

大连理工大学

Dalian University of Technology

摘 要:

软件行业属于高科技产业范畴，整体薪资水平一直处于行业领先。据《2019 年众达朴信软件行业薪酬福利调研报告》数据显示，行业主管层级员工一二类城市年度总现金中位值分别为 20 万和 12 万元.该层级技术研发类员工薪资水平较高，预计 2020 年技术序列涨薪幅度将超过 8%；另外，近些年随着移动互联网、第三方支付、通讯、手机游戏和电子商务的迅速发展，软件行业人才流动率提高，2021 年整体涨幅仍会在 2%以上，研发技术岗会在 10%以上。

知识、技术密集是软件企业的主要特点之一，从业人员主要以本科学历及以上者居多。据某招聘网站 2019-2020 年软件行业薪资数据显示，行业技术类研发毕业生起薪每年均有 3-4%的增幅。一般而言，软件企业更倾向于从重点院校获得优质毕业生，给出的薪资待遇往往更高。近年来，随着 IT 技术的发展，高科技人才稀缺仍然严重，行业知名企业给出的年薪标准已经透明化，使得毕业生对自己的薪资预期往往很高。企业人力资源管理者一方面对于不断提高的人工成本要有应变策略，另一方面也要关注人均效能的提升，才能有效应对这种趋势。

基于调查数据研究软件行业薪资水平的影响因素，可帮助求职者预估薪资情况，也可为软件公司制定薪资方案提供理论依据。数据包括某软件公司员工的工作年限、工作岗位、学历以及薪资。首先对数据进行适当处理，把工作岗位和学历等定性数据进行定量化。之后建立多元线性回归模型刻画工作年限、岗位、学历对薪资的具体影响作用。最后利用 matlab 软件求解模型，该模型通过了数据检验且拟合优度较高。

关键词: 多元线性回归模型；软件行业；薪资变化；matlab；薪资预测

目录

1 引言

软件是新一代信息技术产业的灵魂，“软件定义”是信息革命的重要标志和显著特征，随着新一轮技术创新引领产业新变革，技术创新进入新一轮加速期，云计算、大数据、物联网、移动互联网、人工智能、虚拟现实等新一代信息技术快速演进，单点技术和单一产品的创新正加速向多技术融合互动的系统化、集成化创新转变，创新周期大幅度缩短，硬件、软件、服务等核心技术体系加重重构，新业态、新模式快速涌现，全球软件和信息技术服务业正在步入加速创新、加速迭代、群体突破的爆发期，加速网络化、平台化、服务化、智能化、生态化演进。

根据工业和信息化部统计数据，2018 年，全国软件和信息技术服务业规模以上企业 3.78 万家，累计完成软件业务收入 63061 亿元。2018 年，我国软件和信息技术服务业运行态势良好，收入和效益保持较快增长，吸纳就业人数稳步增加；产业向高质量方向发展步伐加快，结构持续调整优化，新的增长点不断涌现，服务和支撑两个强国建设能力显著增强，正在成为数字经济发展、智慧社会演进的重要驱动力量。

随着中国软件业的迅猛发展，软件工程师的薪资也“水涨船高”，这是吸引大量毕生进入软件行业的主要动力。[1]本文调查了各大招聘网上软件行业的招聘需求和薪资水平，并获取了某软件公司员工的具体薪资及相应的工作年限、学历、工作岗位等数据。基于实际调查数据，利用数学方法，建立了多元线性回归模型，并通过显著性检验，得到了软件行业员工的薪资与工作年限、学历、工作岗位等因素的依赖关系。[2]

2 数据处理与分析

文章的数据来自于国内某招聘网站的的招聘信息，共计 46154 条有关于计算机行业的招聘信息，每一条信息包括：工作岗位、工作地点、公司名称、公司规模、公司类型、学历要求、薪资、福利待遇字段。因为数据很多都是文本结构，所以对数据进行定量化分析，对公司规模、公司类型、学历要求、薪资福利进行了转换，并对薪资进行了平均值的计算。

对公司类型进行了划分，将其分为了'外资'，'民营'，'国企'，'合资'，'上市'，'事业'以及未知七类分别标记为 1-7；将公司规模划分为'少于 50 人'，'50-150 人'，'150-500 人'，'500-1000 人'，'1000-5000 人'，'5000-10000 人'，'10000 人以上'7 种分别标记为 1-7 号；将学历要求划分为"中技","中专","高中","大专","本科","硕士","博士"，'未知'8 种编号为 1-8；关于工作福利，通过将获取的文本进行分类，以福利待遇的多少来划分，标记为福利待遇的个数，如表 1 所示。

表 1 数据描述

	Company_type	Company_size	Graduation	Jobwelf
Count	45503	45503	45503	45503
Mean	2.38	2.27	3.59	5.31
Std	1.31	2.21	1.19	3.02
Min	1	0	0	1
Max	7	7	7	13

Pearson 相关系数是用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。文章利用量化的数据求解 Pearson 相关系数，初步判断得到的数据是否对最后的预测有利，初步的结果如表二。

表 2 各字段相关性分析

	Company_type	Company_size	Graduation	Jobwelf	Salary
Company_type	1.00	0.03	0.03	0.02	0.06
Company_size	0.03	1.00	0.10	-0.18	-0.01
Graduation	0.03	0.10	1.00	-0.03	0.07
Jobwelf	0.02	-0.18	-0.03	1.00	0.03
Salary	0.06	-0.01	0.07	0.03	1.00

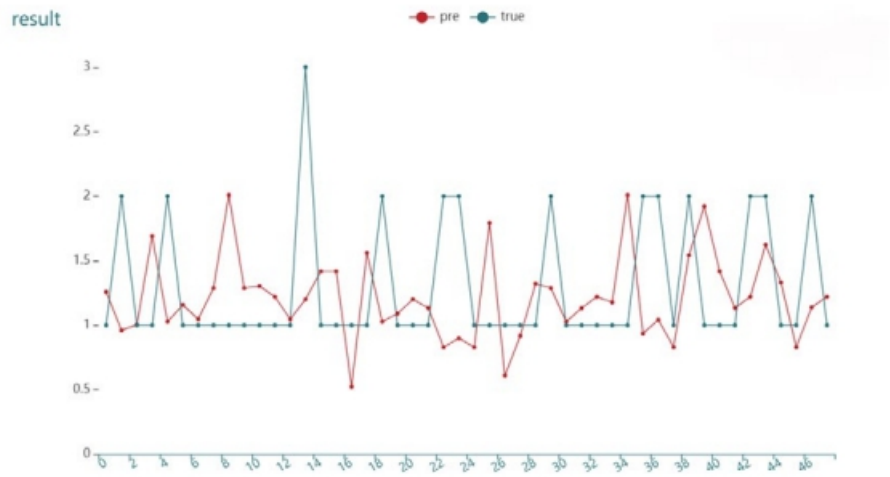
3 预测方法与调整

利用 Pytorch 平台搭建的三层网络结构，通过量化的公司类别、学历要求、工作福利作为输入特征，三层的神经网络[3]，进行预测工资薪酬，以千为单位，文章进行了 300 次迭代训练，模型的 loss 值趋于稳定。但是预测结果趋于一致在 8k 左右徘徊，效果不是很好，预测结果不够理想的原因可能有数据量化处理比较简单和数据来源比较单一，未进行严格的筛选。在此篇文章暂不列出。

虽然通过全连接的神经网络为达到比较好的预测效果，因为有很大一部分原因是数据处理的过程没能做到标准化，以及数据来源不一定全部都是有效的，需要做进一步筛选。为了进一步继续实验得到更多有用信息，利用提取到的现有数据，进行接下来的实验，稍微改变一下模型分类预测策略，将薪酬划分为 10000 以下标记为 1，10000-20000 标记为 2，20000-50000 标记为 3，50000 以上标记为 4 进行分类预测。

在 1980 年 Ohlson 第一个将逻辑回归方法引入财务危机预警领域，他选择了 1970～1976 年间破产的 105 家公司和 2058 家非破产公司组成
的配对样本，分析了样本公司在破产概率区间上的分布以及两类错误和分割点之间的关系，发现公司规模、资本结构、业绩和当前的融资能力进行财务危机的预测准确率达到 96.12%。逻辑回归分析方法使财务预警得到了重大改进，克服了传统判别分析中的许多问题，包括变量属于正态分布的假设以及破产和非破产企业具有同一协方差矩阵的假设。文章想利用逻辑回归的方法，根据公司规模、公司类型、学历要求和福利待遇字段进行逻辑回归训练，预测对应薪酬的等级，如前文所划分的 4 类。通过 sklearn 准确率计算得到准确率的值 0.78，预测分析图（如图 1）

图 1 预测结果分析图



4 结论

本文通过搭建爬虫获取某招聘网站 4 万多条招聘信息，经过一系列的数据清洗处理后，通过数据分析及挖掘得到如下一些结论：（1）公司规模对薪酬的影响不大；（2）通过多元逻辑回归可以基本预测薪酬的等级；（3）IT 行业薪酬普遍过高，对于更好预测薪酬需要引入更多的字段信息。

5 总结与展望

文章利用公司规模、公司类型、学历要求和福利待遇字段信息想预测薪酬，但是因为数据的单一来源，字段的过少，数据处理过程简单等原因没有达到比较理想的结果，后续的研究将引入更多的字段包括，城市，岗位类型，工作经验等更多信息，利用深度学习网络重新搭建模型，实现薪酬的更好预测。

6 参考文献

[1] 刘文远，李少雄，王晓敏，等．大数据知识发现 [J．燕山大学学报，2014，38(5) : 377-380 .

[2] 加春燕,尚意展,钟宇辉,冯文昊,罗丹.多元线性回归模型在软件行业薪资影响因素分析中的应用.《北京工业职业技术学院学报》- 2019 年 3 期

[3] 梅端 周会会.基于神经网络的城镇在岗职工平均工资预测.《河北工程技术高等专科学校学报》.2016 年第 1 期 36-38,共 3 页