个人求职推荐

PERSONAL RESUME

工资预测

PERSONAL RESUME

第 9 组

小组成员：崔　杰 42023053

黄安琪 42127028

陈　丽 42127006

赵俊晖 42138019

一、背景介绍与问题定义

随着大数据时代的来临，数据分析已经成为人们作出决策的关键环节，从而推动了数据分析等相关工作的需求迅速增长。在这样的背景下，数据分析相关职位的招聘市场也呈现出蓬勃的发展态势。据招聘网站数据显示，目前共有 1936 家公司发布了 2913 条数据分析相关工作的招聘信息。

然而，求职者们往往面临众多的选择和挑战。他们需要综合考虑职位名称、公司地区、公司类型、薪资水平、工作经验要求、公司规模以及福利待遇等多个因素，以做出最佳的职业选择。因此，为了帮助求职者更好地理解和选择数据分析相关职位，本文究旨在深入挖掘和分析这些招聘数据，为求职者提供更加科学及可行性高的建议。

首先，本文将预测薪资水平作为一项重要的研究内容。通过分析招聘数据中的多方面信息，以了解当前数据分析职位的市场薪酬水平，以及不同类型公司和地区的薪资差异。这有助于求职者更好地了解市场行情，从而在求职过程中设定合理的薪资期望。

其次，本文将关注哪些类型的公司或领域具有更高的工资水平。通过对公司类型、所在地区、公司规模和公司领域等特征进行深入分析，我们可以发现哪些类型的工作机会更有可能提供高水平的薪资。这对于求职者来说具有重要的指导意义，可以帮助他们更准确地定位自己的职业发展方向。

此外，为了更好地满足求职者的个性化需求，本文将尝试建立一套推荐系统。该系统将根据求职者对工作地点、薪资水平、工作经验和福利待遇等要求，为其推荐最适合的公司或公司领域。通过这种方式，本文旨在帮助求职者更快地找到符合自己需求的职位，提高职业选择的效率和满意度。

总之，本文以深入挖掘和分析招聘网站中的数据分析职位数据为基础，旨在为求职者提供全面、准确的可行性建议。通过建立聚类分析（尝试后所得效果不佳，暂未采用）、主成分分析（尝试后所得效果不佳，暂未采用）、决策树回归模型、线性回归模型、随机森林模型、GBDT 模型，并在处理数据方面做出改进，逐步提升效果，以预测薪资水平、分析不同类型公司和领域的薪资差异以及建立个性化推荐系统，为求职者提供更具体和有针对性的指导，帮助他们更好地理解和选择数据分析职位。

二、数据预处理与特征选择

（一）数据预处理

（I）删除完全重复的记录。原始数据共有 2913 条记录，去掉重复后剩 2595 条数据。（II）由于实习生样本较少，而实习生的薪资远低于入职的工资且计算

方式不同，因此选择直接删除实习生数据，也即删除职位名中含有兼职、实习以及薪资中含有天、时这些关键词的数据。共去除 301 条数据，剩下 2294 条数据。（III）删除工资这一被解释变量为空的数据，仍然剩下 2294 条数据。（IV）对职位名以及福利使用使用 2-5 的词频统计（不包含字符类型，只包含数字和英文），统计出不同长度的前 100 个字，删除频次小于 10 次的词，并进行人工进行语义分析，再进行删除筛选，结果如下图：

图 1 福利关键词

图 2 职位名关键词

（二）变量处理与模型建立

（I）根据“职位名”（原始数据中变量对名称）生成变量高级职称 Senior。从职位名的词频统计表可以看到，其中有很多表示高级的词汇，比如“专员”、“工程师”等，并且我们还发现“算法”、“开发”这两个词汇与“人工智能”等表示工作内容的词汇经常一起出现，因此我们认为“算法”、“开发”两个词也表示更高阶的工作内容。最终我们根据如果职位名中是这类关键词来生成分类变量 senior 表示高级。

（II）根据“职位名”生成新变量工作类型 Jclass。根据词频统计中含有的关键字，对职位内容进行分类，最终筛选出 10 个关键关键词。并统计发现有 40 个职位名共计 57 个样本含有两个及以上分类，选择删除。

（III）根据“薪资”生成新变量年工资 y_Salary。提取薪资变量中的最大最小月工资，并提取月薪份数，出于样本考虑，把月薪份数大于 16 的统一取值为 16 没有月薪份数的处理为 12 薪，最终计算每个工作的最大最小年工资。并且我们认为大多数求职的工资应该是工资范围的中位数，因此为模拟工资的中位数，最终的年工资=（最大年工资-最小年工资）*0.35+最小年工资。并对最终的年工资进行 1%的所谓处理。在运用到模型中时，将工资全部取对数。

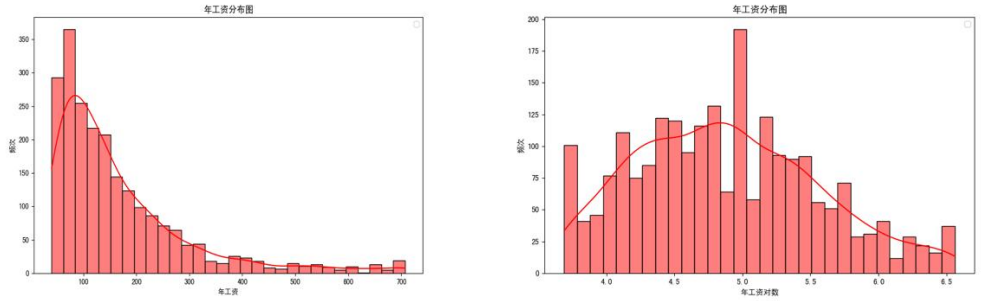

图 3 直方图数据分析

（IV）根据“经验”生成工作经验要求变量 Exp。我们观察原始数据发现，原始数据中工作经验要求分为 6 类，考虑到样本平衡，选择将工作经验要求分为 5 个类别，删除 9 个工作要求为 10 年及以上的样本。

（V）根据“福利”生成股票期权 stack、五险一金 binsurance、补充医疗保险 minsurance、带薪年假 paleave、定期体检(fpe)以及额外福利数 Ebenefit。额外福利数=总的福利数-含有上述 5 项福利的个数。出于样本均衡考虑，将部分取值合并。

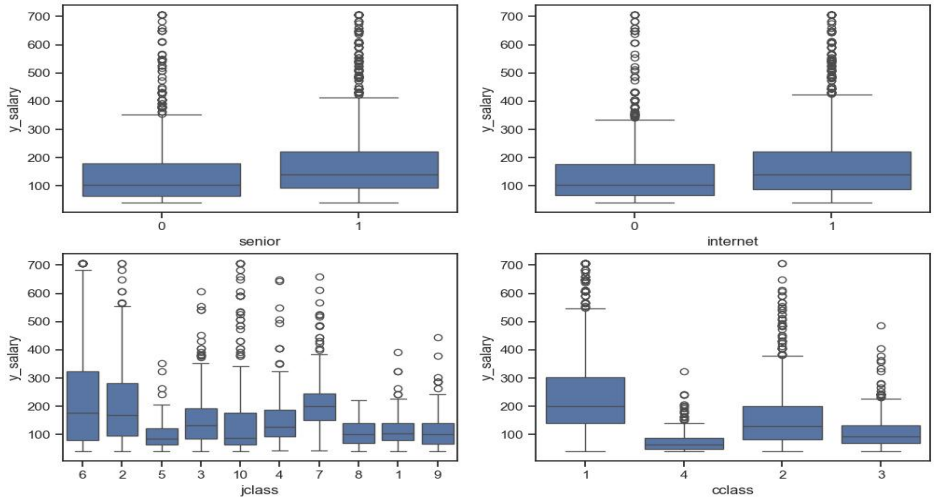

图 4 箱线图数据分析

三、数据介绍

全部特征变量见下表：

表 1 变量说明

变量名	变量类型	变量说明
Senior	分类变量	工作名称中是否含有工程师、高级、专家、专员、研究院、资深、算法、开发这 8 个关键词,含有则取 1，否则取 0
Jclass	分类变量	1.爬虫；2.数据分析；3.后端；4.前端；5.运维；6.人工智能（含机器学习、深度学习、AI、数据挖掘）；7.大数据；8.测试；9.python；10.其他
y_Salary	连续变量	模型当中对工资取对数
Edu	分类变量	0:学历不限；1:大专学历；2：本科；3:硕士及以上
min_Exp	分类变量	原始数据 0 到 1 年工作经验；1 到 3 年工作经验；3 到 5 年工作经验（后续为处理方便，我们只取最低要求工作年限）
CSize	分类变量	公司规模：原始数据分为 6 类，我们将公司规模分为 0-100，100-999，1000-9999，10000 及以上。
Stack	分类变量	工作福利中是否含有股票期权
Binsuracne	分类变量	工作福利中是否含有 5 险一金
Minsurance	分类变量	工作福利中是否含有补充医疗保险
Paleave	分类变量	工作福利中是否含有带薪年假
Fpe	分类变量	工作福利中是否含有定期体检

Ebenefit	连续变量	工作福利中除了上述 5 项福利额外的福利数
Cclass	分类变量	将城市划分为一线城市、新一线城市、二线城市、三线及其他城市
Internet	分类变量	工作领域是否含有关键词：O2O,电子商务，互联网，互联网金融，计算机服务，计算机软件，社交网络，数据服务，信息安全，移动互联网，游戏，含有则取 1，否则取 0

四、决策树回归

（一）决策树回归初步训练

将预处理好的数据中的 80%作为训练集，20%作为测试集。直接带入决策树回归模型进行拟合。

（二）调参前后决策树回归

本文使用 GridSearchCV（网格搜索）方法进行参数调整，由于是回归问题，选取 scoring=”r2”为评分标准。调参的对象为决策树最大深度（max_depth）、内部节点再划分所需最小样本数（min_samples_split）、叶子节点最少样本数（min_samples_leaf）、结点划分时考虑的最大特征数（max_features）根据 best_params_的最佳参数输出结果，在最佳范围做分析。

考虑到网格搜索的参数范围和步长设置可能不够精细，可能导致搜索过程中错过了一些潜在的优良参数组合，优先采取手动调参，尝试根据经验找到较优设置。所得结果为均方误差 0.187，可决系数 0.591。之后进行网格搜索寻找最优参数组合，所得结果为均方误差 0.194，可决系数 0.576。网格搜索后的结果比搜索前的不理想，可能是因为该模型的分类变量较多，维度较高，而决策树只有一颗树，导致性能不佳。

表 2 网格搜索前后调参取值情况

参数名	**参数含义**	**手动调参取值**	**网格搜索调参取值**
max_depth	最大深度	8	10
min_samples_split	内部节点再划分所需最小样本数	2	10
min_samples_leaf	叶子节点最少样本数	4	3
max_features	结点划分时考虑的最大特征数	10	auto
备注	random_state=0，且其余均为默认		

表 3 网格搜索前后调参结果对比

前后	**均方误差 MSE**	**可决系数 R2**

手动调参	0.187	0.591
网格搜索调参	0.194	0.576

现展示网格搜索调参前后的决策树画图，图 4-1 为网格搜索调参前，图 4-2 为网格搜索调参后，图 4-3 为基于调参后结果所得防止过拟合的剪枝后的图。可见，在网格搜索调参前所得决策树划分没有调参后细致，但均存在过拟合的情况，在剪枝后，该求职问题在决策树环境下结果仍有待提升，同时我小组将采用其他模型进行对比。

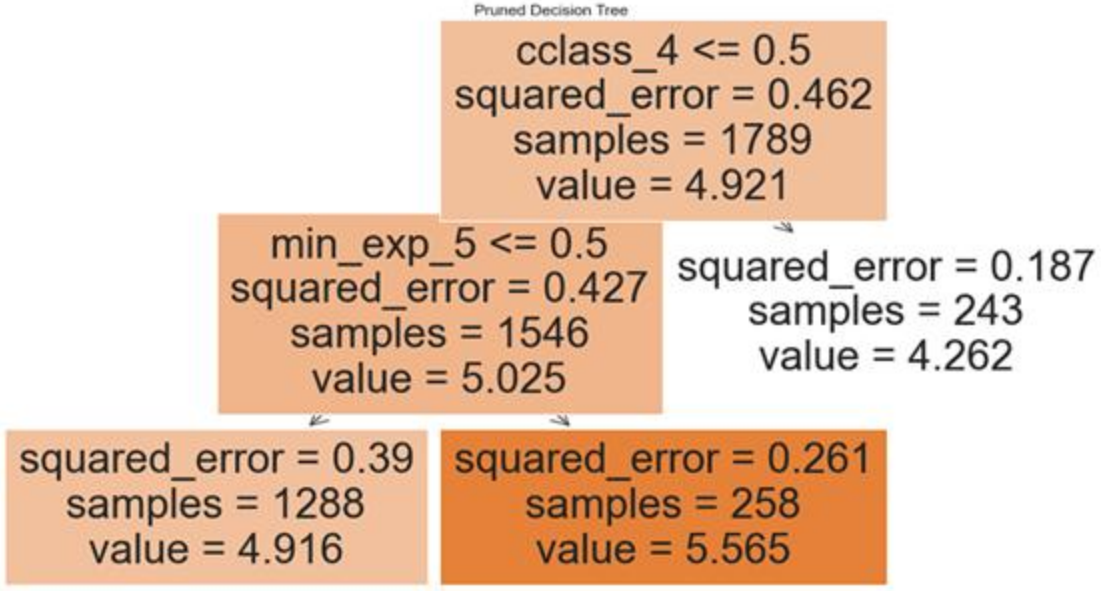

图 5 调参后结果所得防止过拟合的剪枝后的图

模型的评估：由图 6 可见，预测值小于真实值的相对较少，其分布相对均匀，但较为分散，同时仍存在少部分明显偏离 45 度线的异常值。该模型效果有待提升与改进。

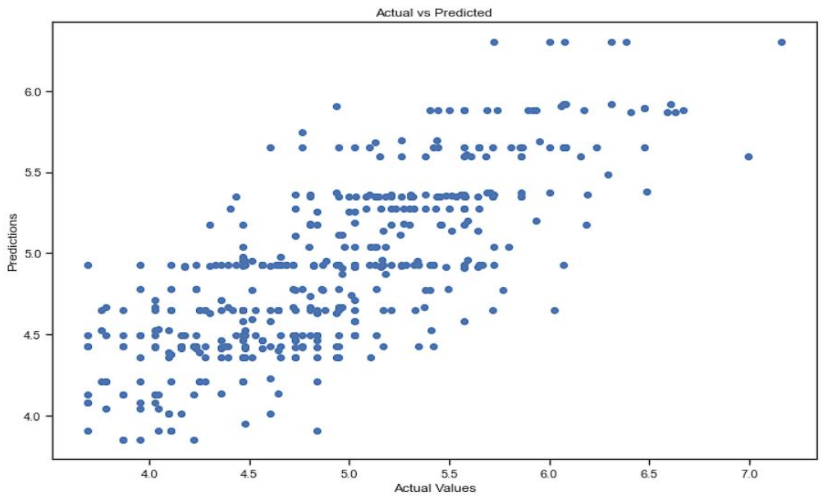

图 6 真实值与预测值散点图

（三）结论分析

在特征重要性方面，由图 7、图 8，网格搜索前后排第一位的都是 cclass_4 城市等级，可见在该模型中，随着城市等级提高，薪资会相应增加；排在最后一位的都是 jclass_7，可见其影响程度较小，特征重要性弱；而处于居中位置的为 exp 经验下限，其具有一定影响力，求职者可以以此为依据进行考量。

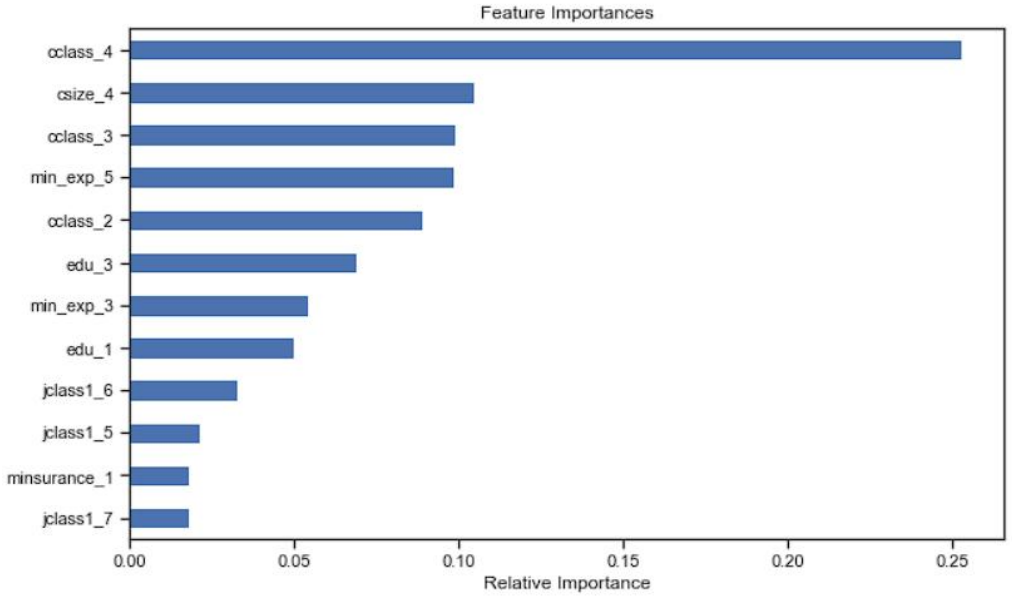

图 7 决策树回归下手动调参各变量重要性

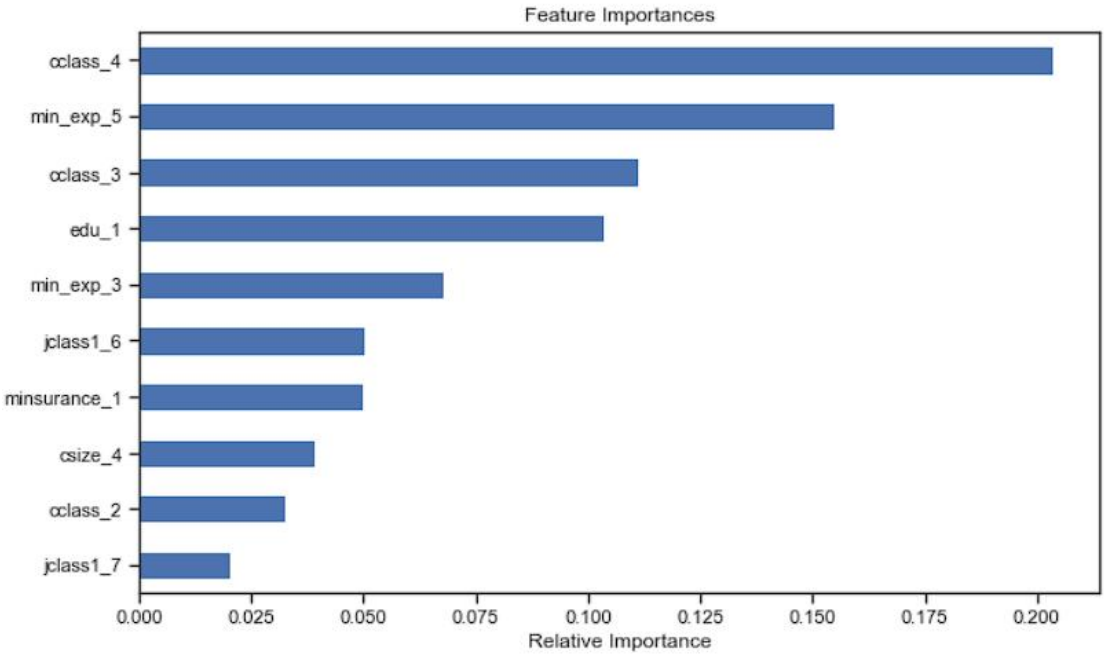

图 8 决策树回归下网格搜索调参各变量重要性

五、线性回归

（一）线性回归模型的参数解释

对年工资取对数以后，选择部分分类变量系数解释如下：这些系数是从线性回归模型中提取出来的，表示各个特征对模型输出的影响程度。

1.**senior_1**：该特征的系数为 0.140，表示含有“高级”词的职位与模型输出之间存在正相关关系。高级职称的工作比普通职称的工作的工资高 14%。

2.**stack_1**：该特征表示股票期权，其系数为 0.123，表示享有股票期权的工作比不享有该福利的工作的工资高 12.3%。

3.**minsurance_1**：该特征表示补充医疗保险，其系数为 0.091，表示享有补充医疗保险的工作比不享有该福利的工作的工资高 9.1%。

4.**jclass_6**：**jclass_6** 表示人工智能相关的工作，其系数值最大，为正的 0.160，意味着相对基准组爬虫工作，人工智能的工作要高 16%。

6.其他变量想城市等级、工作经验要求、文化程度的系数大小及正负都属于正常值。

（二）线性回归模型系数大小与拟合情况总结

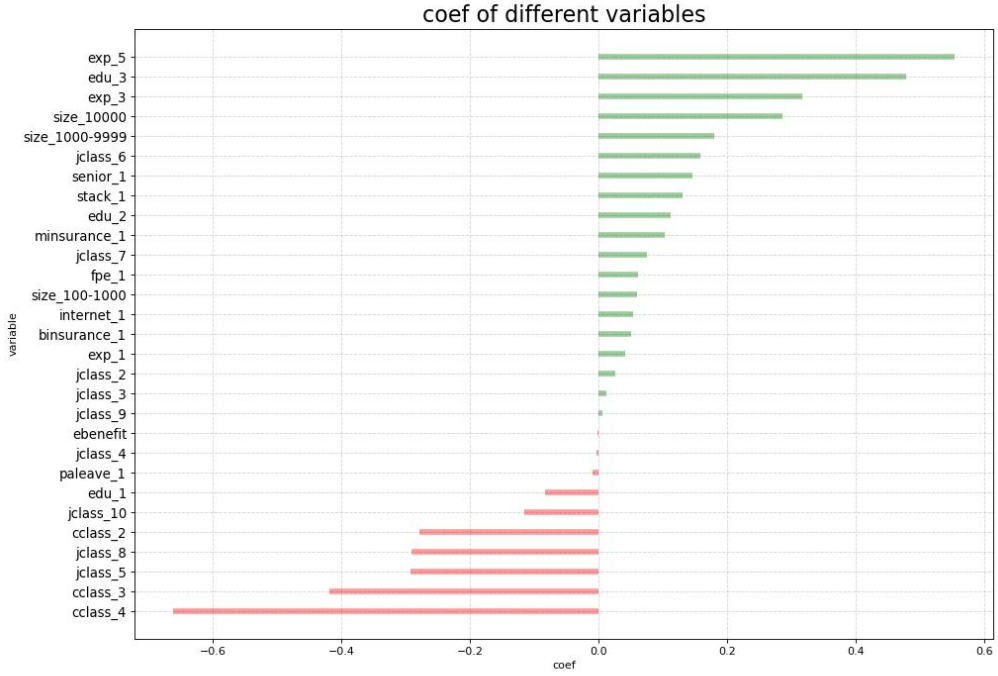

图 9 线性回归相关系数图

模型的评估：由图 8、图 9 可见，用线性回归方法估计的模型，在测试集上的可决系数 R2 为 0.663，均方误差 MSE 为 0.154，均方根误差 RMSE 为 0.393，可见在测试集上模型对薪资的预测效果较佳，模型的拟合程度较好。从真实值与

预测值的对比图来看，由折线图可见，预测值比较贴近真实值，重合度较高，说明预测相对较好，但真实值中仍存在少部分数据偏离较大；由散点图可见，预测值小于真实值的相对较少，其分布相对集中于 45 度线，可以推断其正相关关系较强，但存在极少部分明显偏离 45 度线的值。

线性回归基于一些严格的假设，如误差项的独立性、同方差性等。这些假设在现实世界的数据中可能很难满足，导致模型误设。同时也存在数据线性关系假设：线性回归假设因变量和自变量之间存在线性关系。然而，现实生活中的数据之间的关系可能是非线性的，这使得线性回归在处理这些问题时可能不太有效。所以用线性回归模型预测薪资还有待提升，并不是最优方法。

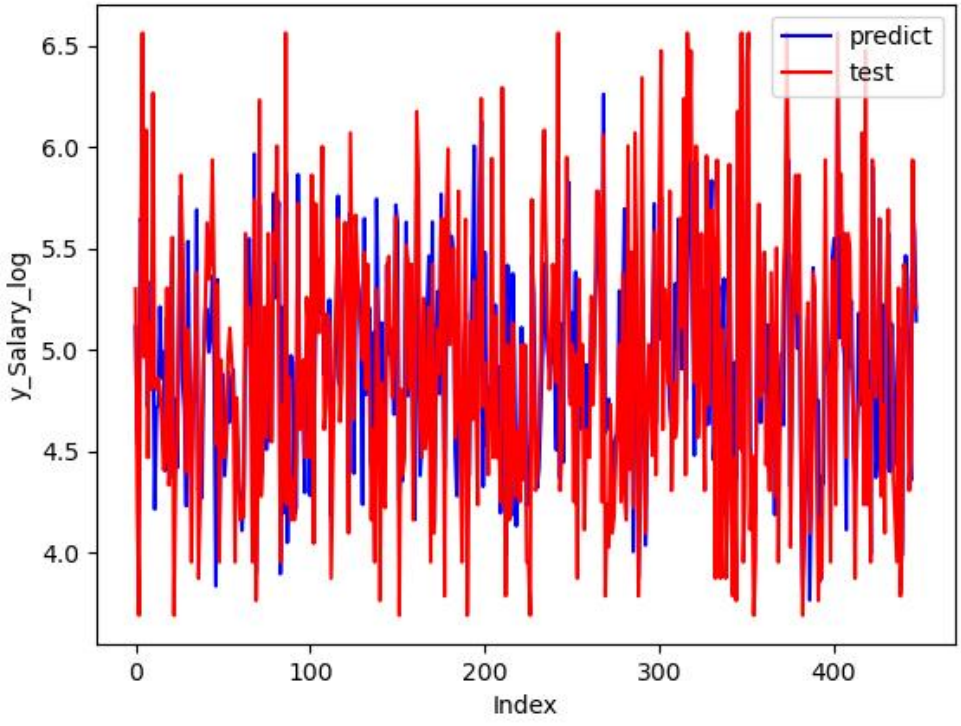

图 10 真实值与预测值折线图

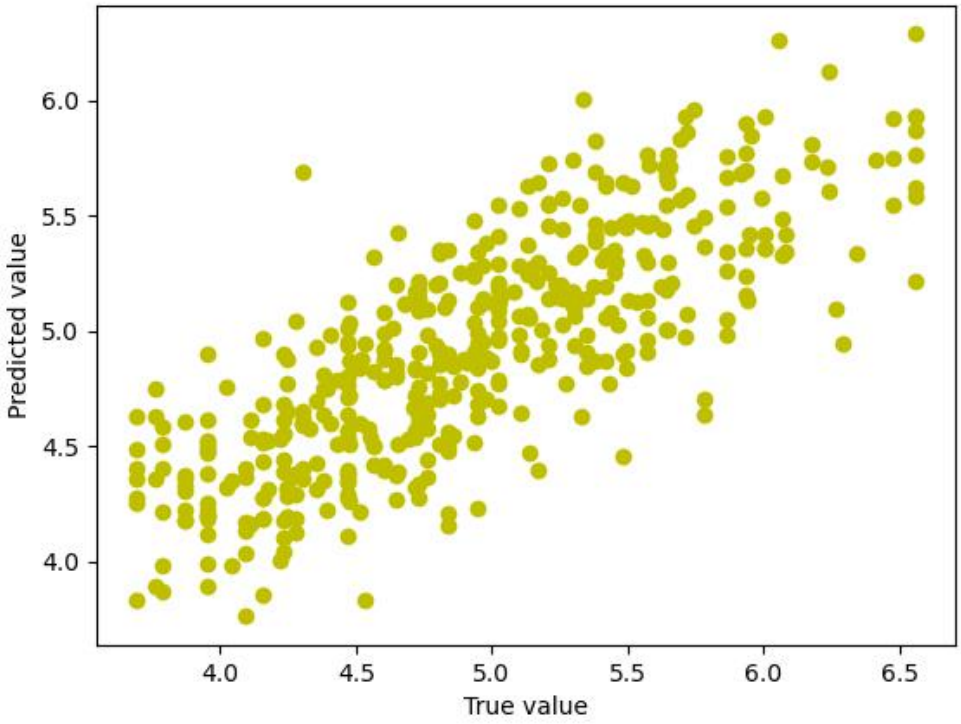

图 11 真实值与预测值散点分布图

（三）决策树回归模型与线性回归模型对比分析

线性回归模型和决策树回归模型是两种不同的回归方法，它们在处理数据和预测结果方面有不同的优势和局限性。线性回归模型效果更加直观，解释性相对较强，所得结果较佳，与决策树回归所得结果对比而言，拟合度更高。但两种回归模型均有待提升，简评二者的差异性原因。

1.**数据特性**：线性回归假设数据之间的关系是线性的，如果数据中的关系是线性的或接近线性的，那么线性回归模型可以很好地拟合数据。而决策树回归不具有这样的假设限制，可能在非线性数据上表现更好。

2.**模型解释性**：线性回归模型相对简单，更容易解释和理解。决策树回归虽然也可以提供直观的解释，但在处理复杂数据时可能不如线性回归直观。

3.**参数调整**：对于决策树回归，参数调整是一个重要的环节，如树的深度、叶节点最小样本数等。如果参数设置不合适，可能会导致模型过拟合或欠拟合。而线性回归模型的参数较少，调整相对简单。

4.**异常值和噪声**：线性回归对异常值和噪声较为敏感，而决策树回归在这方面可能表现更好。

5.**特征选择**：决策树回归模型具有特征选择的能力，可以自动选择对预测结果重要的特征。而线性回归需要手动选择特征或使用其他特征选择方法。

综上所述，对于求职问题这一项目，线性回归模型在性能上优于决策树回归模型。

六、随机森林

（一）随机森林拟合

最后将处理好的数据中的 80%划分为训练集，20%作为测试集，随机种子设置为 0 进行回归拟合。

第一步均使用默认参数进行随机森林模型回归：

表 4 随机森林默认参数

参数名	参数解释	值
max_depth	最大深度	None
max_features	考虑的最大特征数	sqrt
n_estimators	森林中决策树个数	100
min_samples_leaf	叶子节点最少样本数	1
min_samples_split	内部节点再划分所需最小样本数	2
max_leaf_nodes	最大叶子节点数	None

拟合后的结果用散点图表示为

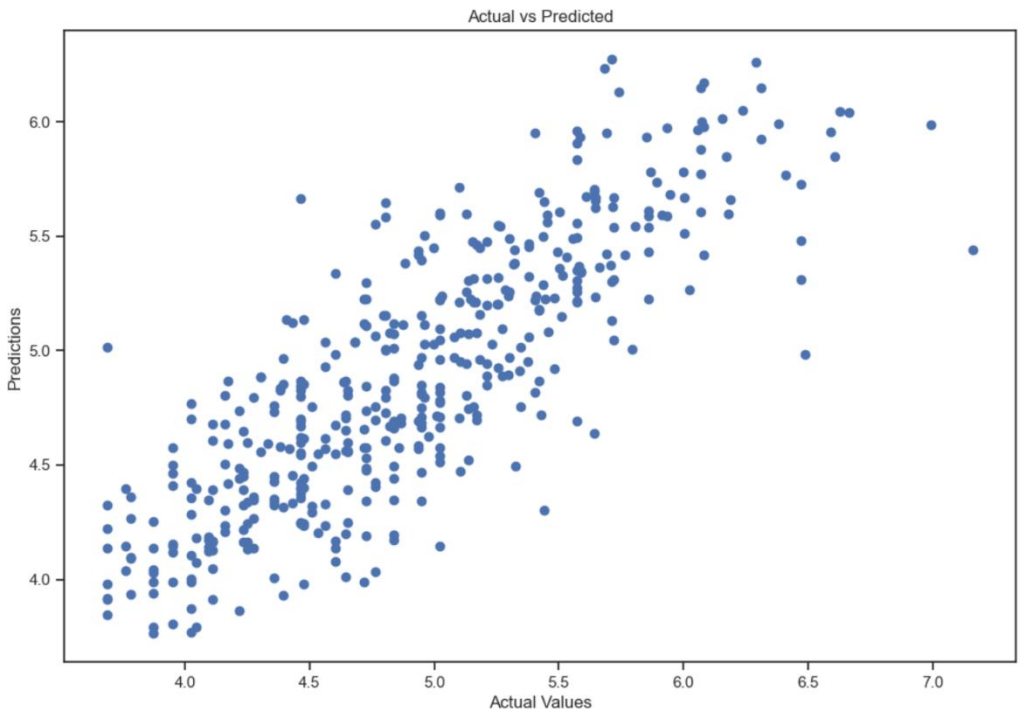

图 12 随机森林默认参数真实值与预测值散点分布图

本文选取了计算袋外可决系数（OOB）、均方误差（MSE）、可决系数

（R-squared）三个值进行回归效果的评估。可以看出预测效果较好，袋外可决系数达到了 0.6126，均方误差控制在 0.1463，相关程度为 0.6799。

（三）调参后随机森林拟合

为了使模型的回归效果更好，本文考虑使用对 GridSearchCV（网格搜索）方法进行参数调整，同时为了防止过拟合的现象，结合了五折交叉验证法进行了参数调整。主要的调整的参数范围如下：

表 5 网格搜索法的参数网格

参数名	参数解释	范围
ccp_alpha	后剪枝调整参数	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5]
n_estimators	森林中决策树个数	[50, 100, 200, 300, 350]
max_depth	最大深度	[12, 15, 20, 22, 24, 26]
min_samples_leaf	叶子节点最少样本数	[1, 2, 4]
max_features	考虑的最大特征数	['auto', 'sqrt', 'log2']

通过调参拟合，得出的最优参数组合为：后剪枝调整参数：0.0，最大深度：22，考虑的最大特征数：auto，叶子节点最少样本数：4，森林中决策树个数：100。相对于默认的参数值而言，最大深度从 10 增加到了 22，考虑的最大特征数从开方后的数量变成了让算法根据数据集的特点自动决定每棵树应该考虑多少个特征，叶子节点最少样本数由 1 变成了 4，其余均保持不变。在此最优的参数组合下，拟合情况如下：

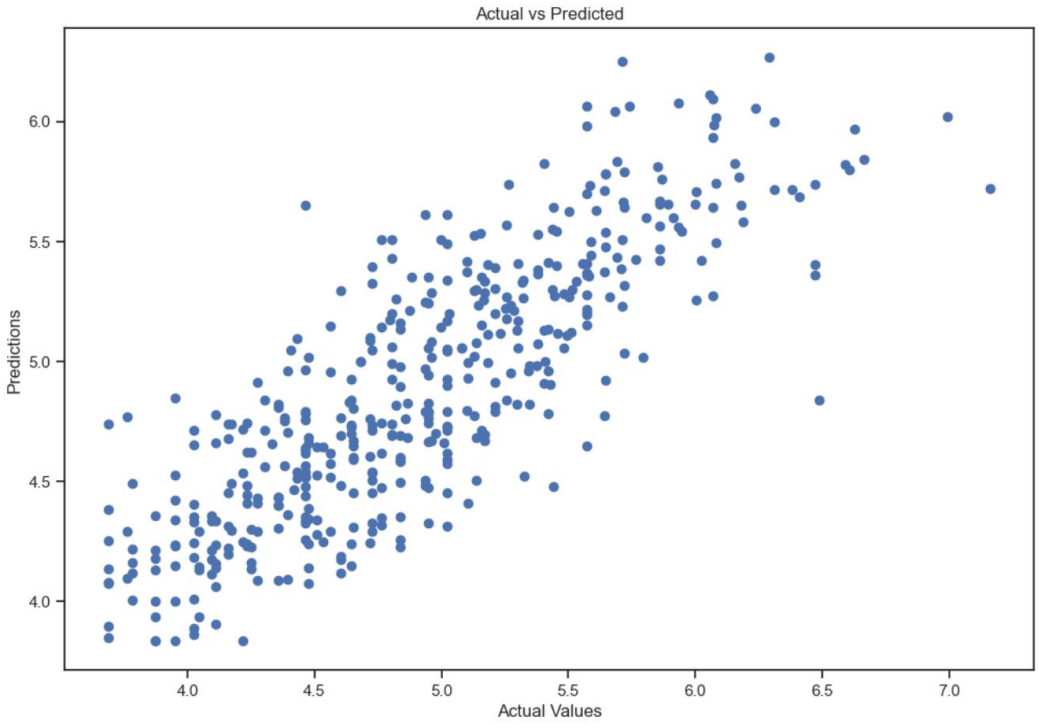

图 13 随机森林最优参数组合真实值与预测值散点分布图

对比默认时候的散点图情况可以发现整体更向 45°先汇集，但是情况不明显。结合 3 个评估标准，即此时袋外可决系数为 0.6280，均方误差控制在 0.1448，相关程度为 0.6833 可以看出，经过调参后该案例中袋外可决系数发生了略微的增长，均方误差降低，而可决系数提高了一些，整体的预测效果也较好。说明该模型在原基础上发生了正向的改进，但整体的改进效果并不明显。

（四）结论分析

在本次调参中，分别选择了排名在前十重要的虚拟变量进行可视化：

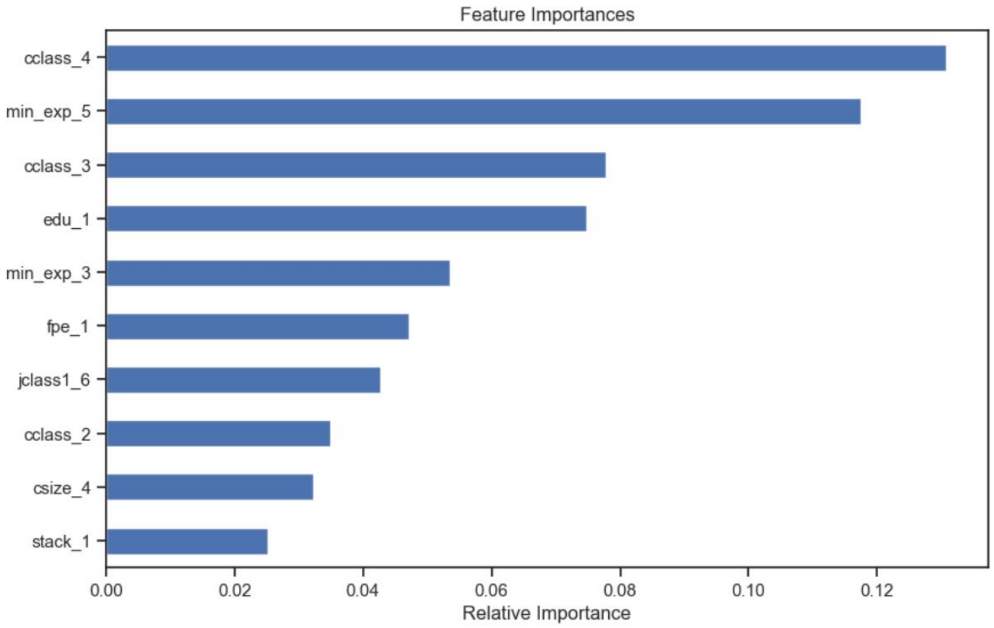

图 14 随机森林原始参数下的特征重要性排序（前十）

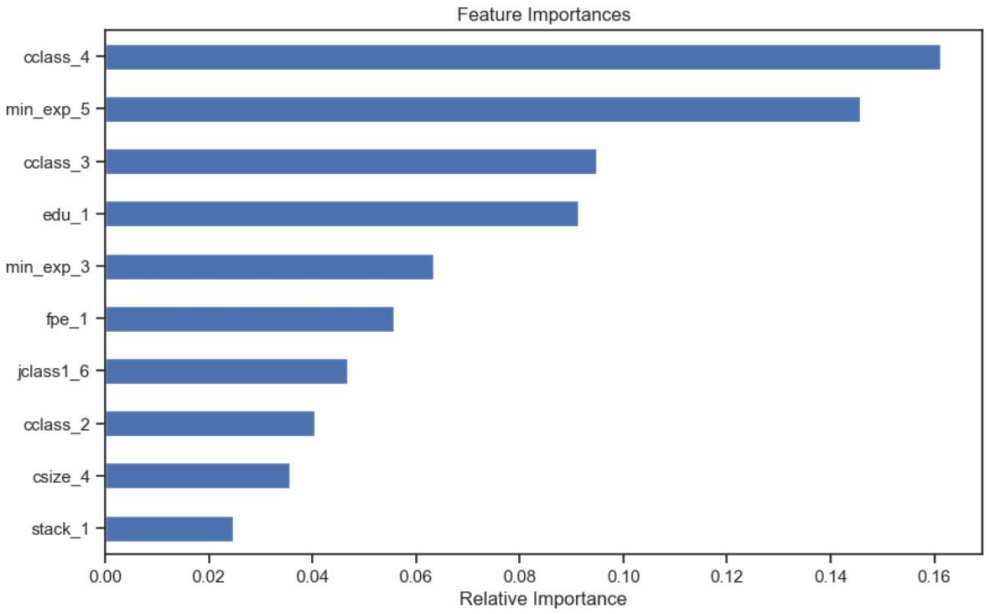

图 15 随机森林最优参数组合下的特征重要性排序（前十）

选取了重要性排名前 5，中间 5 个，以及最后 5 个参数和重要性数值进行分析：

从总体分布上可以看出，虚拟变量特征重要性排序有所区别，但整体上变量的排序变化不大。在变量中首尾特征的重要程度相对减小，而排名在中间的特征重要性增加了。

对于工资水平而言，城市的等级、工作经验和受教育水平的对其影响较大，而各种年薪津贴福利等对工资水平的影响相对较小，一些不常用不重要的比如体检等服力待遇对于工资而言影响很小。另外，不同的职位关键词等对工资的影响区别较大。

七、GBDT

GBDT 利用损失函数的负梯度在当前模型的值作为回归问题提升树算法中的残差的近似值，拟合一个回归树。相较于 AdaBoost 提升法，GBDT 在抗噪声表现方面更强。

（一）模型调参：基于 TPE 的 Hyperopt 调参方法

按照导包等准备，读取数据，搜索超参，定义目标函数，带 CV 方法交叉验证，定义超参空间，执行超参搜索，获取最优参数，绘制搜索过程，训练模型的步骤实现调参。

选用了基于贝叶斯优化的 Hyperopt 的超参数寻参方法，可以规避网格搜索人为设定的具体数值，以设定参数的寻参范围，分别以连续值、离散值进行寻参，提高了参数搜索精度。

原始参数空间：以连续值不断带入目标函数当中，参数空间范围内全部匹配，缺点是速度慢：

n_estimators: [1, 200]

learning_rate: [0.1, 1]

max_features: “sqrt”, “auto”, ”log2”

subsample: [0.1, 0.9]

loss: “ls”, “lad”, “huber”, “quantile

max_depth: [1,60]

min_impurity_decrease: [1,20]

根据初步拟合得出的参数结果，对参数空间进一步放缩，加快了拟合效率，提升了寻参速度：

n_estimators: [75, 325]

learning_rate: [0.01, 0.45]

max_features: “sqrt”, “auto”, “log2”

subsample: [0.2, 0.65]

loss: “ls”, “lad”, “huber”, “quantile

max_depth: [24, 69]

min_impurity_decrease: [2,18]

（二）模型预测

将处理好的数据的 80%作为训练集，20%用于测试集，同时选择随机森林模型中，经过网格搜索的调参数据作为基学习器，有效改进 GBDT 模型剪枝过粗的问题。将经过 Hyperopt 调整过后的参数带入模型之后得到的 R2 为 0.706，MSE 为 0.134，基本不存在过拟合的现象。

（三）结论分析

排名前十的特征重要性：

调整数据集之前：

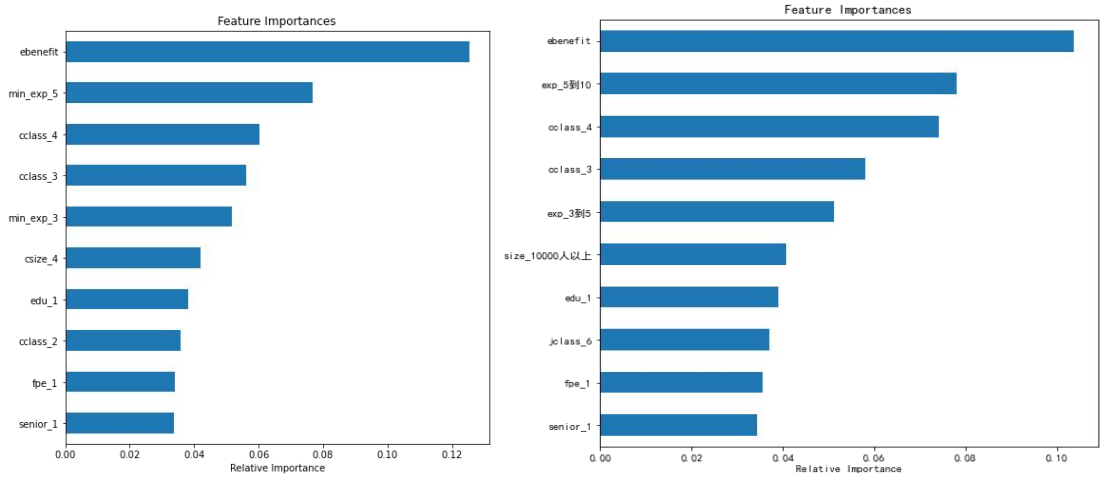

图 16 调整数据集前后重要性

八、模型比较与总结建议

（一）原始数据集调整

在数据集的处理方面，展示时我们的数据集中月薪份数与工作福利中的“年终奖挂钩”，并且没有关注到额外福利数不同取值的样本样本数相差较大这一情况，我们根据上述两个问题对数据集进行了调整。

（二）模型比较

各个模型的 R2、MSE 如下图所示，

本案例选取的变量众多，数据量大，GBDT 回归效果最好，可能是基于下面的原

因：

1.随机森林是一个集成多个决策树的强学习器，对其进行了精剪枝，有效抗过拟合；

2.随机森林抗噪声能力强，能处理特征较多的高维数据

3.GBDT 采用了随机森林作为基学习器，能处理包含大量虚拟变量的高维数据；

4.GBDT 能够降低对于异常点的敏感程度，在噪音比较多的数据集上表现更加优异。

调整数据集之前：

表 6 调整数据集前效果

回归效果	决策树回归	线性回归	随机森林	GBDT
MSE	0.194	0.154	0.145	0.134
R2	0.576	0.663	0.683	0.706

调整数据集之后：

表 7 调整数据集后效果

回归效果	决策树回归	线性回归	随机森林	GBDT
MSE	0.221	0.177	0.171	0.158
R2	0.529	0.623	0.635	0.664

（三）总结建议

在线性回归、决策树、随机森林、GBDT 四种模型当中，通过比较 R2 得分和 MSE 大小，本文选择 GBDT 作为模型来预测薪资水平。

对比不同模型排名前十的特征重要性，从外部环境，个人能力，公司待遇三个方面对薪资水平的预测进行解释。

1.外部环境方面，主要包括城市发展水平 cclass 这一变量。

城市发展水平 cclass，城市等级越高，薪资水平越高，可能是由于工作所在城市的经济发展状况的基础水平决定。所有城市相较于一线城市的工资要低，其他城市与一线城市的收入差距达到 7.4%的水平。

2.个人能力方面，主要包括工作经验 exp 和学历要求 edu 两个变量。

工作经验 exp，工作经验越多，薪资水平越高。可能是拥有较多的工作经验，能让求职者更快匹配岗位工作要求，减少入职后的技能学习培训时间，在岗位上创造更高价值。求职者工作经验丰富，有利于在薪资水平谈判的中争取到更多权益。

学历要求 edu，学历水平越高，薪资水平越高。本科学历以及硕士及上学历

工作的工资明显高于大专以下及学历不限工作的工资。有趣的是，edu_1 对工资的影响比 edu_3， edu_2 还分别高约 1.1%和 1.6%，可能是由于较早的毕业进入工作，能让公司较早的进行业内培训，与此同时积累了更多的经验导致的薪资水平差异。

3.公司待遇方面，主要包括 jclass，公司规模 size，福利水平 ebenefit

岗位类型 jclass 对薪资水平有一定影响，不同岗位类型工资水平差异大，可能是由于某些特定的岗位能创造更高的价值，数据挖掘（jclass_6）类型的工资较高，后端（jclass_3）、前端（jclass_4）等“苦力”活的工资较低。而岗位名称中是否为高级岗位 senior，公司类型是否为信息互联网行业 internet 则对薪酬水平影响较小。

公司规模 size 对薪资水平方面影响较小，10000 人以上公司规模比 1000-9999 人的公司规模的薪资水平高出约 2%。

福利水平 ebenefit 对薪资水平的影响最大，达到 10.4%的影响水平。在薪资水平中，定期体检（fpe）对薪资水平的影响达到约 3%的水平，股票期权（stack），补充医疗保险（minsurance）对薪资水平的影响达到约 2%的水平，侧面反映对待员工福利待遇比较重视的公司，给出的薪资水平越高。

结合上述影响薪资水平的变量设置，我们建议求职者优先考虑公司的福利待遇水平，结合公司所在城市的发展水平和公司规模，寻找一线城市，大规模公司的就职机会，并在岗位类别中，从事与人工智能相关的工作。努力学习，提升自身学历，积累工作经验，有利于获得更高的工资。

小组分工：

崔　杰 ：数据预处理与特征筛选，进行变量说明，统筹安排组内工作，组内代码审核与反馈，PPT 制作与汇总，报告攥写与修改。

黄安琪 ：背景介绍与问题定义，试运行聚类分析，决策树回归建模，线性回归建模，PPT 制作，报告攥写。

陈　丽 ：随机森林工资回归与调参，试运行二次调参，工作城市的随机森林分类，PPT 制作与修改，报告攥写与汇总修改。

赵俊晖 ：GBDT 模型搭建，试运行 AdaBoost 模型，PPT 制作，模型比较与总结建议，报告攥写。