AI报告质量提升研究——问题一

> 请结合数据挖掘、机器学习等方法解决如下问题：
> **问题一**：AI输出报告质量与企业输入属性之间的预测分析
> 要求1：从企业输入属性中，选择与AI输出报告质量相关的属性
> 要求2：从企业输入属性中选择合适的属性与AI输出报告质量进行预测分析
> **本次分析的目标是基于企业输入的多维数据，找出与AI输出报告评分相关的关键属性，并利用这些属性进行预测分析，提升AI生成报告的质量。**

一、数据预处理

1. 输出报告评分与评价的映射关系

- **不合格**：评分在50到55之间
- **合格**：评分在60到65之间
- **良好**：评分在70到78之间
- **优秀**：评分为80

评分和评价之间有明确的映射关系，在分析时使用评分即可充分代表报告的质量评价，所以我在后续分析中主要使用**评分**作为质量评价的指标。

2. 数据完整度的计算

为了更好地衡量企业输入数据的质量，我添加了一个新的指标：“**数据完整度**”，该指标通过统计每个企业的11个属性中非空值的数量来表示。该列可以反映企业输入信息的详细程度，并可能与AI输出报告的评分存在相关性。

3. 空缺值处理

由于部分属性存在缺失值，我对不同类型的数据进行了相应的处理：

- **数值型变量**：如“企业人数”、“月均薪酬”等数值列，使用**中位数**进行填充，以减少缺失值带来的偏差。
- **类别型变量**：如“岗位名称”、“所属部门”等文本列，使用“未知”进行填充。
- **二进制处理**：将“其他”列处理为二进制形式，标记是否存在额外的企业属性。

4. 转换特定列

- **年营业收入水平**：通过去掉括号内容并将万、亿单位转换为数字，处理为可用于计算的数值格式。
- **成立时间**：将成立时间的区间或符号处理为中间值，保证数据的一致性。
- **企业人数、月均薪酬、直接下属人数**：这些列通过正则表达式处理，将不规则数据格式转换为中间值。

5. 保存处理后的数据

最终处理后的数据已保存为新的Excel文件，文件名为 `处理后的数据_完整度.xlsx`，其中包含完整的预处理结果。该文件可用于后续的特征选择和建模分析。

序号	岗位名称	所属部门	企业类型	所属行业	企业人数	发展阶段	成立时间	年营业收入水平	月均薪酬	直接下属人数	其他	数据完整度	输出报告的评分	输出报告的评价	扣分点
1	开发工程师	IT部	港澳台商投资企业	煤炭开采和洗选业	34	未知	8	7.5e+06	1500	2	0	7	65	合格	不要有第一或者第二人称；无法确定横向部门是否描述准确；无法判断所开发的软件/系统是是否与实际一致。
2	数据分析	it	有限责任公司	计算机	74	未知	8	7.5e+06	6500	0	0	7	70	良好	无法确定实际是分析内部数据还是外部数据，有可能结果完全跑偏
3	水利工程项目经理	未知	集体企业	水利管理业	1500	成长期（3-10年，扩张市场，增加产品）	8	1.5e+07	9000	9	0	7	75	良好	内容可以更详尽
4	审计主管	审计部	国有企业	石油和天然气开采业	250	成长期（3-10年，扩张市场，增加产品）	8	7.5e+06	9000	4	1	10	60	合格	对项目审计提得偏少
5	互联网产品项目经理	未知	股份有限公司	互联网和相关服务	150	生存期（1-3年，市场适应，现金流管理）	8	4e+06	9000	9	0	8	75	良好	nan

二、特征选择

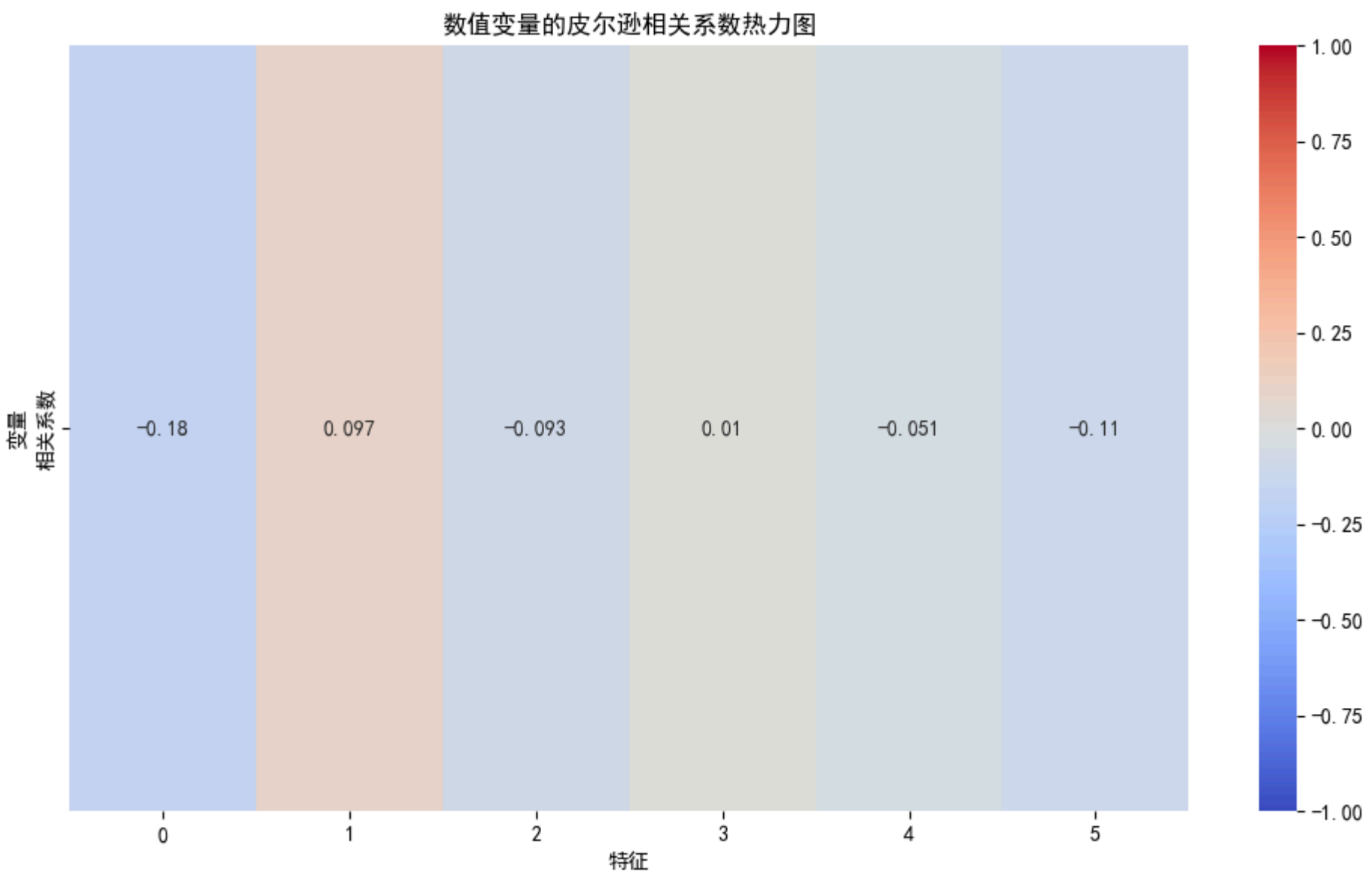
1. 数值变量选择

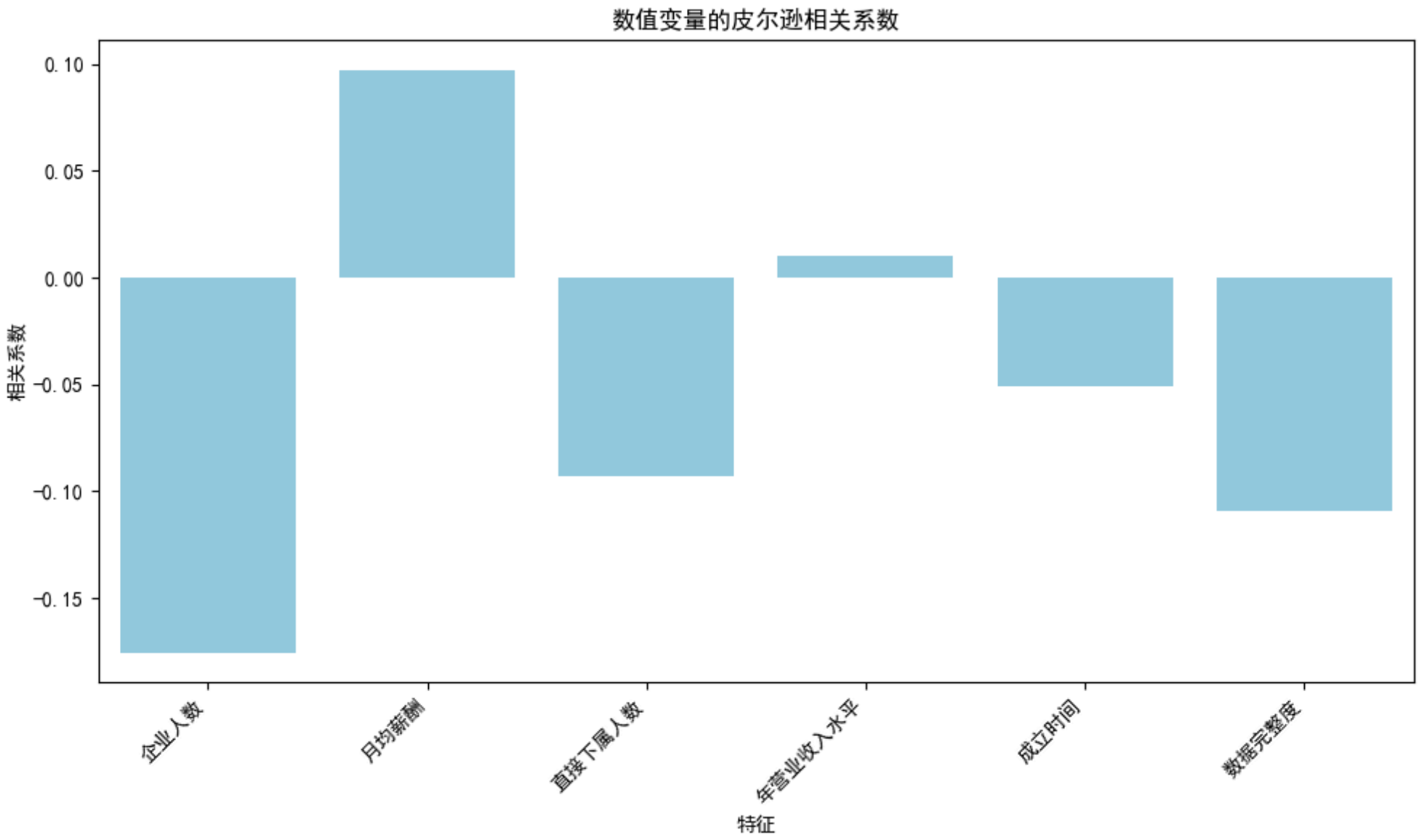
首先，通过**皮尔逊相关系数**分析企业的数值型属性与输出报告评分之间的关系。这些数值型变量包括企业人数、月均薪酬、直接下属人数、年营业收入水平、成立时间以及数据完整度。

可视化解释：

- **热力图**展示了数值变量与输出报告评分的皮尔逊相关系数。相关系数的范围从 -1 到 1，表示变量之间的线性关系。颜色越接近红色，表示正相关越强；越接近蓝色，表示负相关越强。图中显示了不同数值变量与评分的相关性：
  - **企业人数**：相关系数为 **-0.175**，颜色呈现较浅的蓝色，表明企业规模与报告评分存在一定的负相关性。企业规模越大，AI输出报告的质量评分可能会略有下降。
  - **月均薪酬**：相关系数为**0.097**，呈现淡淡的红色，说明月均薪酬与评分存在轻微的正相关性，薪酬水平越高，AI报告的评分可能会稍微提高。
  - **数据完整度**：相关系数为 **-0.109**，颜色偏蓝，显示数据完整度与评分之间的负相关关系。这意味着输入数据越不完整，AI报告质量评分可能越低。
- **条形图**以更直观的方式展示了每个数值变量的相关系数。从图中可以看到，企业人数和数据完整度的相关性较为显著，虽然它们都是负相关，而月均薪酬与评分的正相关性则相对较弱。条形图帮助我们进一步明确了哪些变量与评分有正负相关性。

虽然相关系数值相对较低，但在实际应用中，这些数值变量依然对报告质量有一定影响。因此，我选择了**企业人数、月均薪酬**和**数据完整度**作为数值特征，继续分析它们对AI输出报告评分的影响。





皮尔逊相关系数结果（数值变量）：

特征	相关系数
企业人数	-0.175439
月均薪酬	0.0972891
直接下属人数	-0.0930694
年营业收入水平	0.0103919
成立时间	-0.0508796
数据完整度	-0.109099

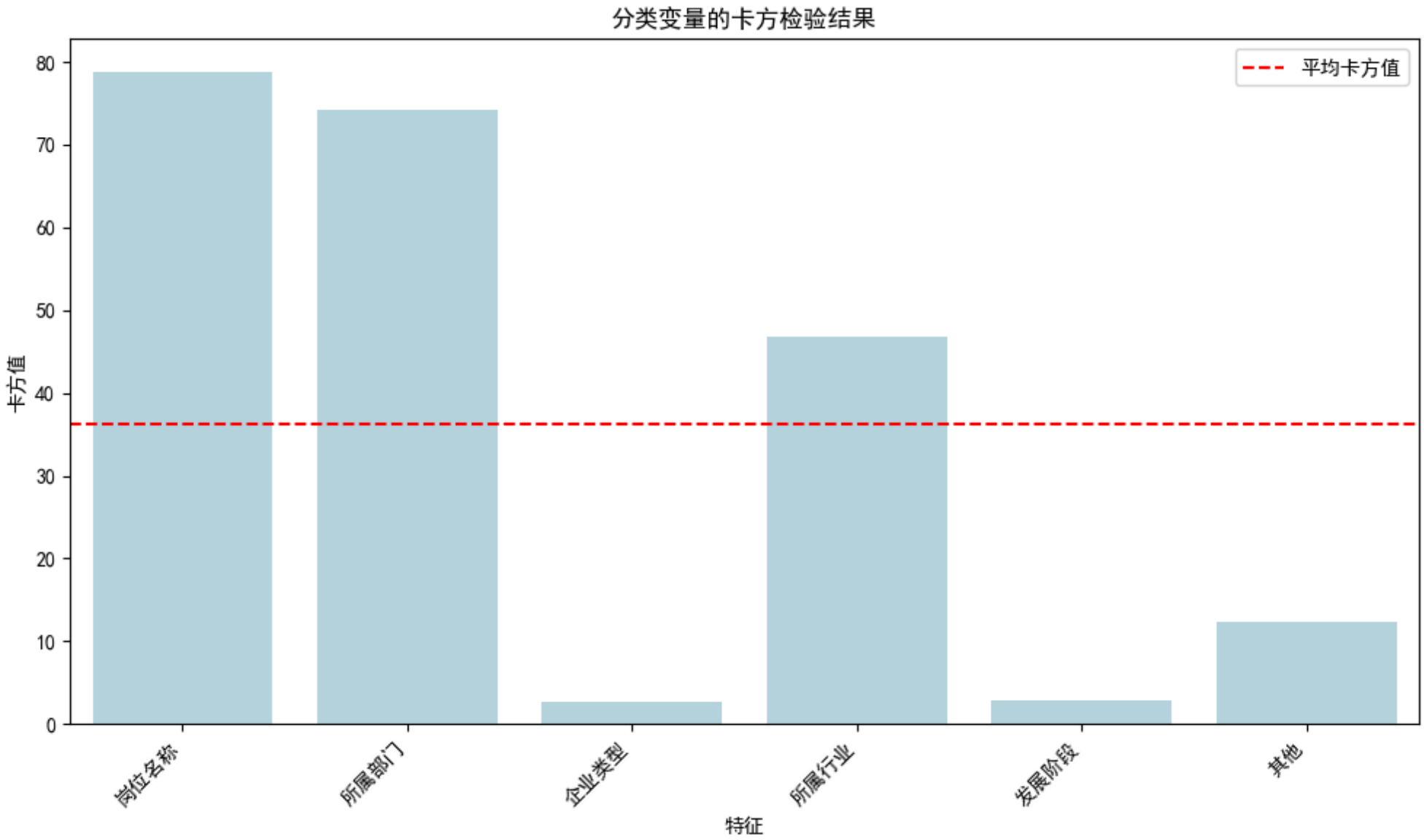
2. 分类变量选择

接下来，我使用**卡方检验**分析了企业的分类变量，包括岗位名称、所属部门、企业类型、所属行业、发展阶段和其他。

可视化解释：

- **条形图**（第三张图）展示了分类变量的卡方值。卡方值越大，表示该变量与AI输出报告评分的相关性越强。图中显示：
  - **岗位名称**和**所属部门**的卡方值最高，分别为**78.79**和**74.26**，说明这些变量对报告评分的影响显著。不同的岗位和部门对AI生成报告的要求和标准不同，因此评分也会有较大差异。
  - **所属行业**的卡方值为**46.79**，显示行业类别对AI报告评分也有显著影响。不同的行业对报告的详细程度和准确性有不同的需求，进而影响评分。
- 红色虚线标示了所有变量的**平均卡方值**。位于红线之上的变量（如岗位名称、所属部门、所属行业）显示出与评分更为显著的相关性，而位于红线之下的变量（如企业类型、发展阶段）对评分的影响较小。

通过卡方检验分析，我选择了**岗位名称、所属部门和所属行业**作为分类特征，重点分析它们对报告评分的影响。这些变量显著影响AI报告质量，并且它们在不同类别之间的评分差异较大。



卡方检验结果（分类变量）：

特征	卡方值	p值
岗位名称	78.7943	2.42575e-14
所属部门	74.2558	2.03073e-13
企业类型	2.66774	0.913946
所属行业	46.7941	6.12271e-08
发展阶段	2.94077	0.890427
其他	12.3648	0.0891839

3. 特征选择的逻辑与结论

基于上述的皮尔逊相关系数和卡方检验分析，最终选择了以下特征：

- **数值变量**：企业人数、月均薪酬、数据完整度。
- **分类变量**：岗位名称、所属部门、所属行业。

但从皮尔逊相关系数和卡方检验的结果来看，这些特征与AI输出报告评分的相关性并不强。

数值变量的相关性分析

皮尔逊相关系数衡量数值变量与评分之间的线性相关性，取值范围为 -1 到 1：

- **|相关系数| > 0.7**：强相关
- **|相关系数| 0.3-0.7**：中等相关
- **|相关系数| < 0.3**：弱相关

在我们的分析中，所有数值变量的相关系数均低于0.2，属于弱相关范围，表明它们对评分的线性影响较小。

分类变量的相关性分析

通过卡方检验衡量分类变量与评分的关联，卡方值越高，表示相关性越强。虽然岗位名称和所属部门的卡方值较高（78.79 和 74.26），但整体来看，分类变量的相关性也不是很强，说明这些变量对评分的影响有限。

三、预测分析

在本次分析中，我选择了**随机森林**模型来预测AI输出报告的评分。随机森林模型具备处理复杂特征和非线性数据的能力，因此被广泛应用于分类和回归任务。然而，在实际应用中，模型的预测效果并不理想，且经过多次调参后，效果依然未能显著提升。

1. 模型效果评估

在使用随机森林模型进行预测时，获得的性能指标如下：

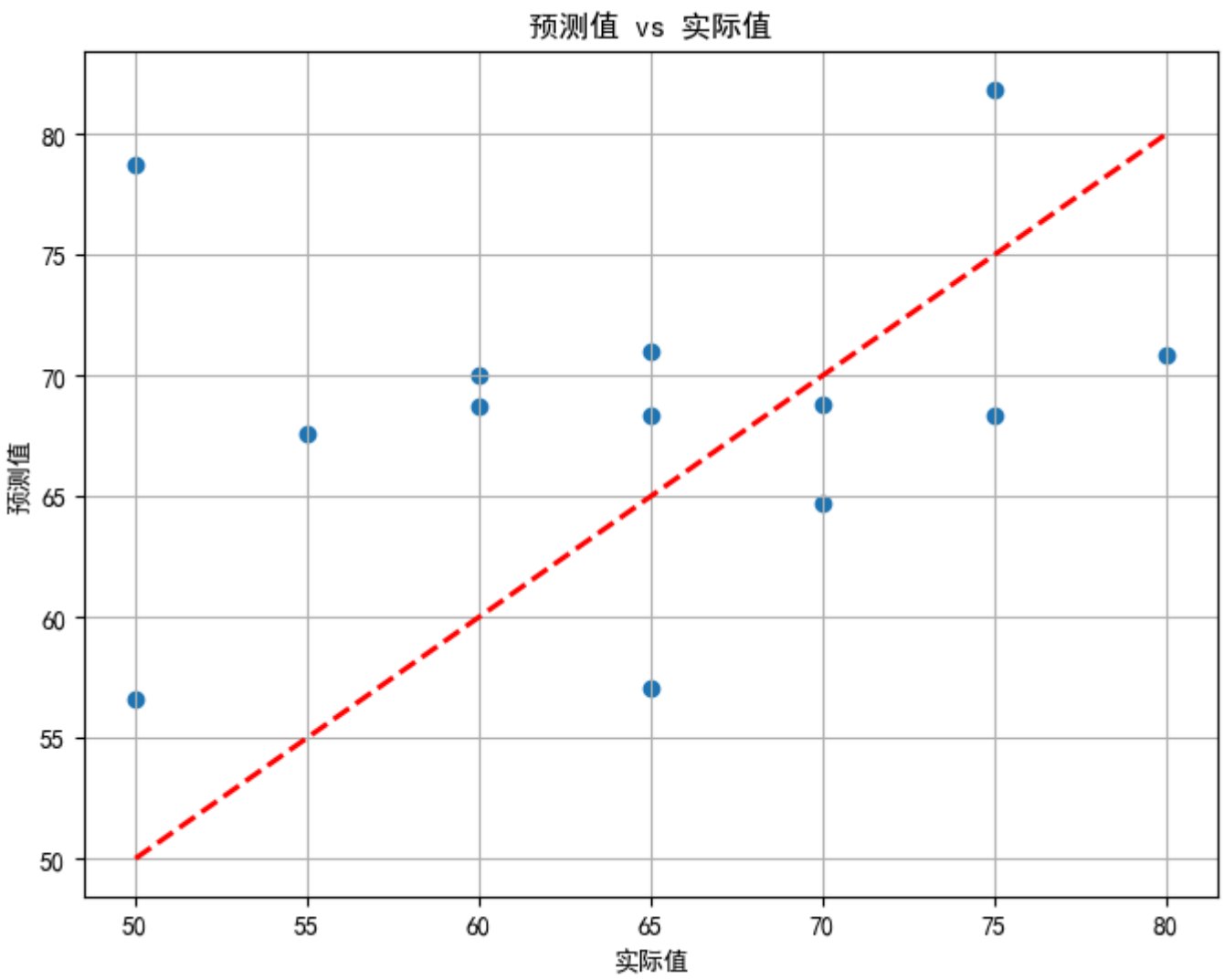
- **均方误差（MSE）**：123.93
- **决定系数（R^2）**：-0.501

从这些结果可以看出，模型的预测误差较大，决定系数为负值，这意味着模型的预测性能甚至比简单的平均模型还差。这表明当前的特征与评分之间的关系较为复杂，随机森林模型未能很好地捕捉到这种关系。

2. 数据可视化分析

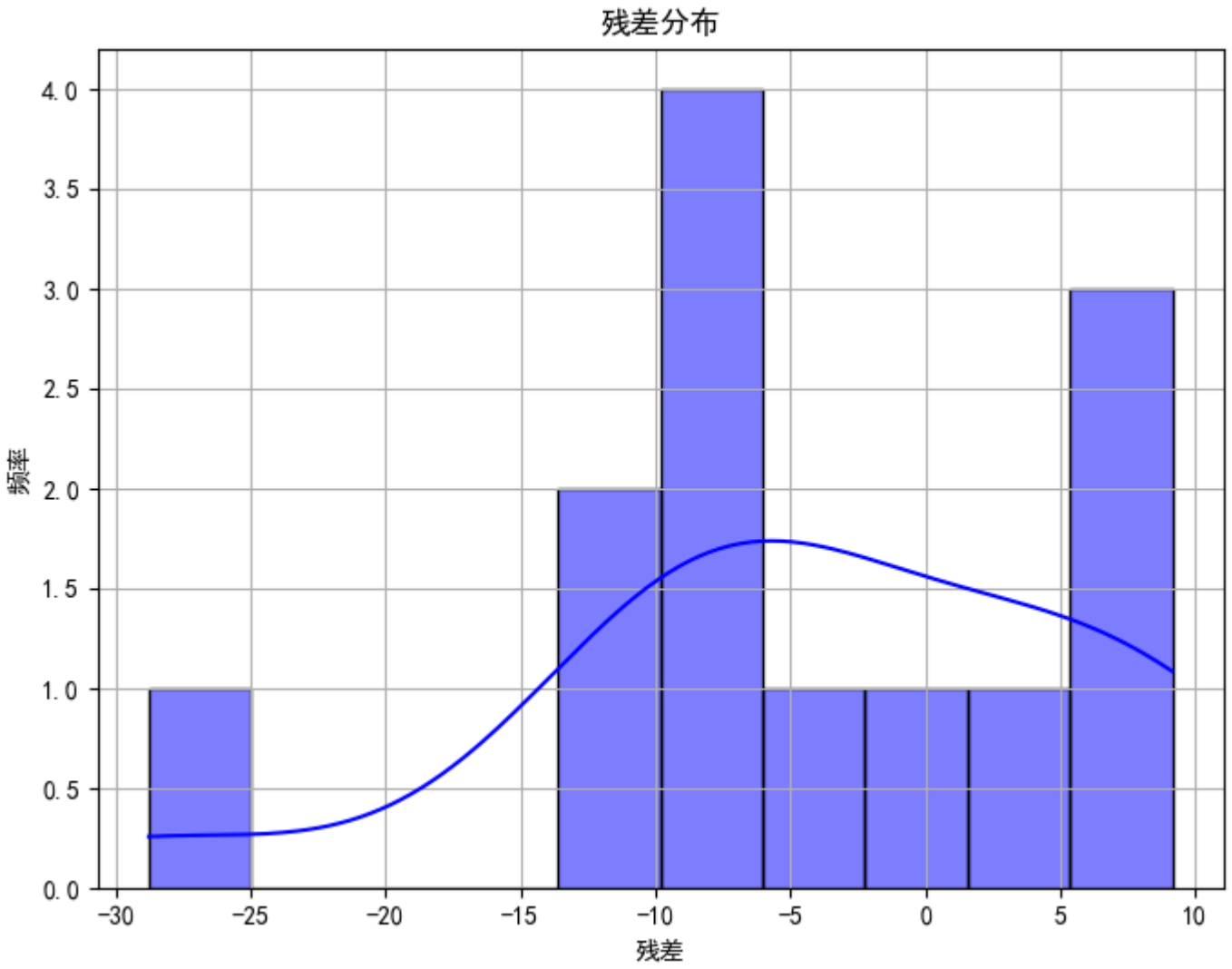
(1) 预测值 vs 实际值散点图

- 图中展示了测试集的实际值与预测值的对比。理想状态下，所有点应沿着红色虚线分布，表示实际值与预测值完全一致。然而，大部分点偏离虚线较远，说明模型在多个数据点上的预测误差较大。



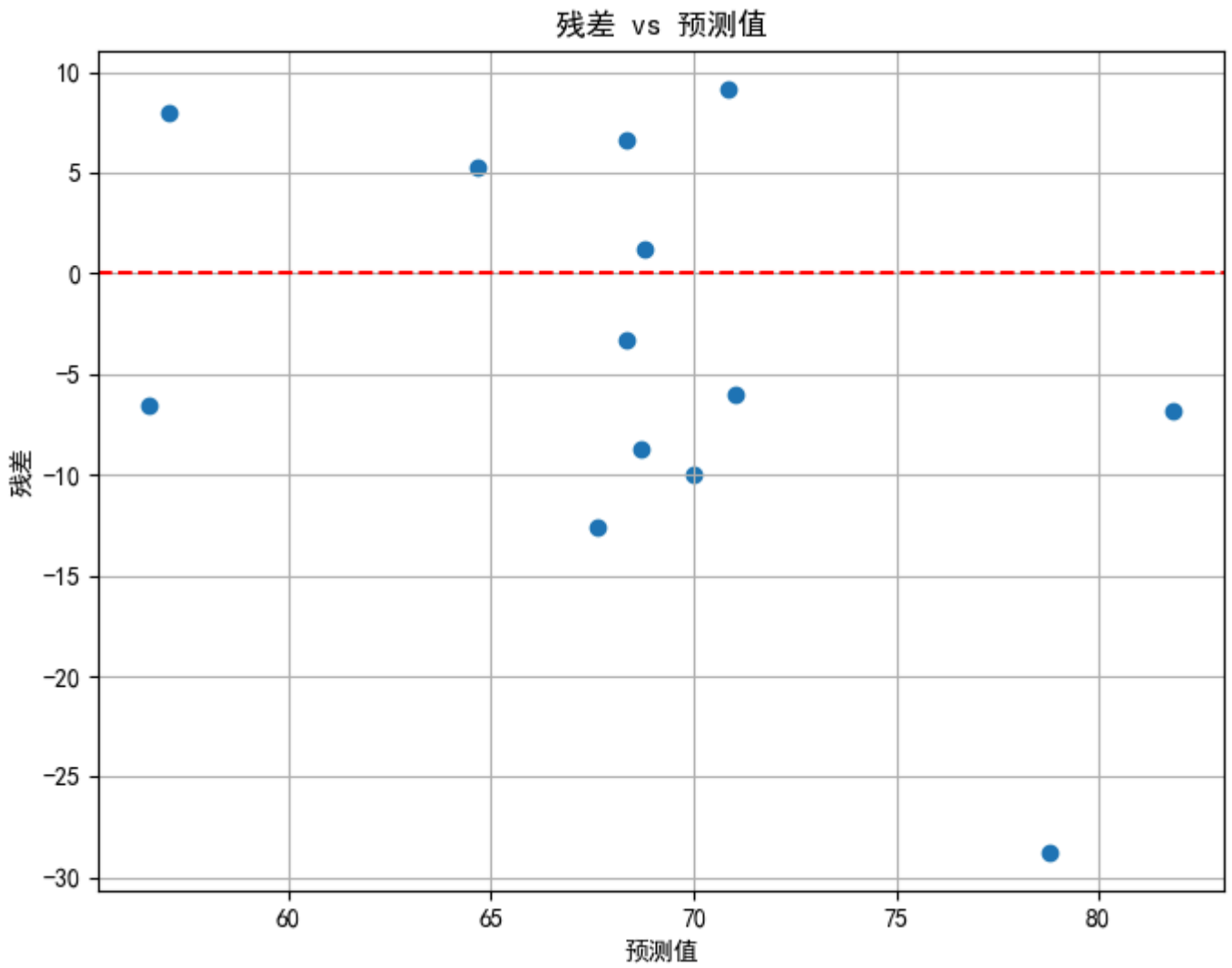
(2) 残差分布图

- 该图展示了模型的残差分布。理想情况下，残差应接近正态分布并集中在0附近。但图中残差呈现出较大偏差，且在-15到5之间的残差值较为集中，表明模型在预测一些数据时误差较大，无法准确反映出实际情况。



(3) 残差与预测值的关系图

- 该图显示了预测值与残差之间的关系。图中显示出较为明显的偏差，特别是在低分和高分段的预测中，残差较大，说明模型无法准确捕捉评分的变化规律。



3. 尝试其他模型

为了验证问题是否出在模型选择上，我还尝试了其他常见的回归模型，但效果依然不理想：

- **线性回归模型**
  - 均方误差（MSE）：117.04
  - 决定系数（R^2）：-0.418
  - 线性回归模型的表现略好于随机森林，但决定系数仍为负值，表明线性回归模型未能很好地拟合数据。

- **岭回归**
  - 均方误差（MSE）：129.33
  - 决定系数（R^2）：-0.567
  - 岭回归模型的结果与线性回归相似，仍未改善模型性能，说明加了正则化后模型表现仍未达到预期。
- **梯度提升决策树**
  - 均方误差（MSE）：117.30
  - 决定系数（R^2）：-0.421
  - 尽管梯度提升决策树具有强大的非线性拟合能力，但在本次分析中，效果也并不理想，说明特征与评分的关系并不容易通过传统回归模型来捕捉。

4. 为什么模型表现不佳?

在尝试了多个模型之后，预测效果仍然不理想，主要原因包括以下几点：

- **特征与目标变量的弱相关性**：正如前面提到的，输入特征（如企业人数、月均薪酬、数据完整度等）与输出评分之间的相关性较弱。由于这些特征难以有效地解释评分的变化，模型无法建立起可靠的预测关系。
- **数据量不足**：当前的数据样本量可能不足以支撑复杂模型的训练，导致模型无法充分学习数据中的模式。特别是在小样本量的情况下，模型容易表现出高偏差或高方差的问题。
- **模型复杂度**：尽管随机森林和梯度提升决策树等复杂模型能够处理高维和非线性数据，但在小数据集上，它们容易出现过拟合或欠拟合的情况。在当前的模型中，决策树可能过于复杂，导致模型对训练数据拟合较好，但在测试数据上表现不佳，从而出现较大的误差和负的决定系数（R^2）。

总结：特征与评分之间的弱相关性，数据样本量不足，以及模型的复杂度共同导致了模型预测效果不理想。要提升模型的性能，未来可能需要增加数据量、优化特征选择，或者尝试更为适合小样本的算法。