招聘岗位薪资预测的最佳模型（新数据集）

1. 问题界定

在课程案例中，我们已经针对一批数据实现了通过公司招聘信息预测岗位薪资。现在又得到了一批新的数据，但是因为数据涵盖的时间、公司范围等有所不同，需要分析人员重新构造预测模型，评估模型效果并确定与薪资相关的招聘信息。本次项目中，请参考课程案例的主要思路，基于新的数据集预测公司岗位薪资。

2. 数据准备

（1）读入新表 job_description.csv 的数据后，请简单描述现有数据的数量与字段情况。

（2）请检查现有数据中是否存在数据缺失、数据异常、数据重复的情况，并做相应的处理。

（3）本次项目的业务问题是预测岗位薪资，现有字段 job_salary 属于文本，请将其转换为连续型变量。同时，字段 company_overview 与 job_info 因为是长文本，无法与其他字段一同作为预测模型的输入，请参考课程案例的思路，选择这两个字段的最佳转换方法。

（4）对于其他输入字段，请通过适当的方法将文字变量转换为模型可处理的数值变量。

字段 company_nature, company_people, job_edu_require, job_exp_require 通过 onehot 编码转换为数值变量，对于转换后的公司概述与岗位概述，同样使用 onehot 编码进行转换。

3. 数据建模与可视化

与课程案例保持一致，本次项目中我们依然比较三种回归模型：MLP、RBF、Lasso 线性回归模型。

（1）请针对三种模型，基于验证集的平均绝对误差，确定模型的最优超参数组合，并比较三个模型的最终效果。

（2）请对 Lasso 模型中非零特征的权重进行解释。

提高题：

数据的预处理方法、模型参数选择等都有可能影响最终模型的效果与对结果的解释。请回顾本次项目中的数据分析流程，尝试修改或添加其他的数据预处理或者参数选择方法，查看分析结果是否能够进一步改善。

提示：

（1）在将文本变量转换为数值变量时统一采用的是 onehot 编码，可以考虑对定类变量进行 onehot 编码，定序变量进行 ordinal 编码；

（2）MLP 模型在 sklearn 工具包中有更多可供修改的参数。