课程概述

您好！欢迎来到数据分析中级课程的第五个案例：基于公司招聘信息的岗位薪资预测。本次案例的数据集比上一个案例增加了一些字段，同时在数据处理的方法上有所突破，具体如下：

1. 内容简介

1）问题界定

基于业务理解和数据理解构造主业务问题。

2）数据准备

除了简单的数据检视与清理工作，重点在于对文本向量化后的高维字段进行改造，使其可以同其他字段一起用于数据建模。

3）数据建模与数据可视化

本案例新增 MLP 模型和 RBF 模型等相关介绍；

同时加上 Lasso 模型共 3 个模型，对其分别进行网格搜索最优超参数，以及相互之间的比较和结果解释；

由于上一个案例中详细地介绍了数据分割和数据重构对模型的影响，因此本案例对该部分内容将进行简化思考。

2. 学习目标

学习完本次案例，你能够达到以下目标：

1）能够熟练掌握异常数据的查找与处理；

2）能够掌握字符串的改造，以及文本向量化后高维数据的改造与转换；

3）了解神经网络模型的基本原理，掌握 MLP 模型和 RBF 模型的使用方法。

业务理解

1. 业务现状描述

在招聘信息中非常重要的一条是岗位薪资，通过对岗位薪资的分析，一方面网站可以基于招聘信息，为招聘公司提供薪资制定方面的建议；另一方面网站也可以基于应聘者的理想薪资，给出个人能力改进方面的指导。

青青招聘网站运营部门的数据分析人员需要根据已有的招聘信息数据，构建岗位薪资预测模型，从而分析公司、岗位信息和岗位薪资的相关性。

2. 识别利益相关群体

本案例中识别利益相关群体：

受项目影响的利益相关群体-----使用青青招聘网站服务的用户；

影响项目的利益相关群体-------系统运营部门人员。

对于系统运营部门而言，他们的需求是获得岗位薪资的影响因素。对于使用网站服务的用户，他们的需求是网站可以提供更好的招聘信息推荐服务。

因此，数据分析人员要在岗位薪资预测模型中得出对公司薪资有影响的因素，同时尽可能地使模型的表现更好。

3. 业务问题构造

本案例的网站为了改进服务，希望通过分析已有的招聘信息数据来预测岗位薪资。

因此，当前的业务问题是：

通过公司招聘信息预测岗位薪资。

数据理解

数据分析中理解数据的第一步往往是了解数据中所有字段的含义，理清数据字段所描述的对象、关系等，然后将数据与业务问题联系起来，方便后续的数据准备和数据建模。

直接查看数据表 job_description.csv 中的字段，发现表中 9 个字段的主要含义如下：

字段	字段含义
company_name	公司名称
company_financing_stage	公司融资情况
company_overview	公司概述
company_people	公司员工数量
job_edu_require	岗位学历要求
job_exp_require	岗位经验要求
job_info	岗位概述
job_name	岗位名称
job_salary	岗位薪资

业务问题需要解决的，就是根据其中 8 个字段来预测岗位薪资 job_salary。