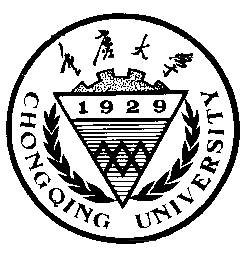
大数据分析实验报告

实验三



学 生：李嘉迪\_20191737

年 级：2019级

专 业：数据科学与大数据技术

重庆大学大数据与软件学院

2022 年 5 月 13 日

## 实验目的：

理解三种或分类模型原理；掌握企业审计数据预处理；掌握分类模型训练、 预测；建立风险分析建模指标；掌握评估方法

## 实验要求：

为了推进企业审计良性发展，减少新企业存在欺诈风险，本实验要求通过企 业审计风险识别相关工作来帮助审计师，建立相应的分类企业审计风险识别模型。 企业审计数据集通过网站下载 https://archive.ics.uci.edu/ml/datasets/Audit+Data， 该数据集是 2015 年至 2016 年间一些公司非机密数据，根据诸如环境状况报告、 利润价值记录等信息，预测公司的欺诈风险可能性，共计 26 个属性。这个风险 主要通过最后一列的 Risk 可以看出，为 1 表示存在风险，为 0 表示不存在风险。

具体实验要求如下：

（1）加载并读取企业审计数据

（2）数据预处理

1)缺失值处理（可以采用零值、均值、分位数等填充）

2)标准化处理 注：一般数据预处理首先看标签列的数据分布和缺失情况，再看特 征列的数据情况

（3）数据集划分

（4）建立三种或以上分类模型训练

（5）对各模型进行效果评估，预测企业是否存在欺诈风险

1) 得出混淆矩阵，并用图表示 2) 评估计算准确率，召回率等 3) 计算出真阳率和假阳率，并绘制 ROC 曲（6）实验分析与总结

## 开发环境：

编程软件：anconda/spyder/pycharm

环境：python3.6 以上、numpy、pandas、sklearn、Jupyter Notebook 等

## 实验结果：

图形用户界面

低可信度描述已自动生成

图表

描述已自动生成

图形用户界面

描述已自动生成

文本

描述已自动生成

## 问题和解决：

得到的结果曲线过于平滑

数据过拟合