Adult 数据挖掘

宋璨

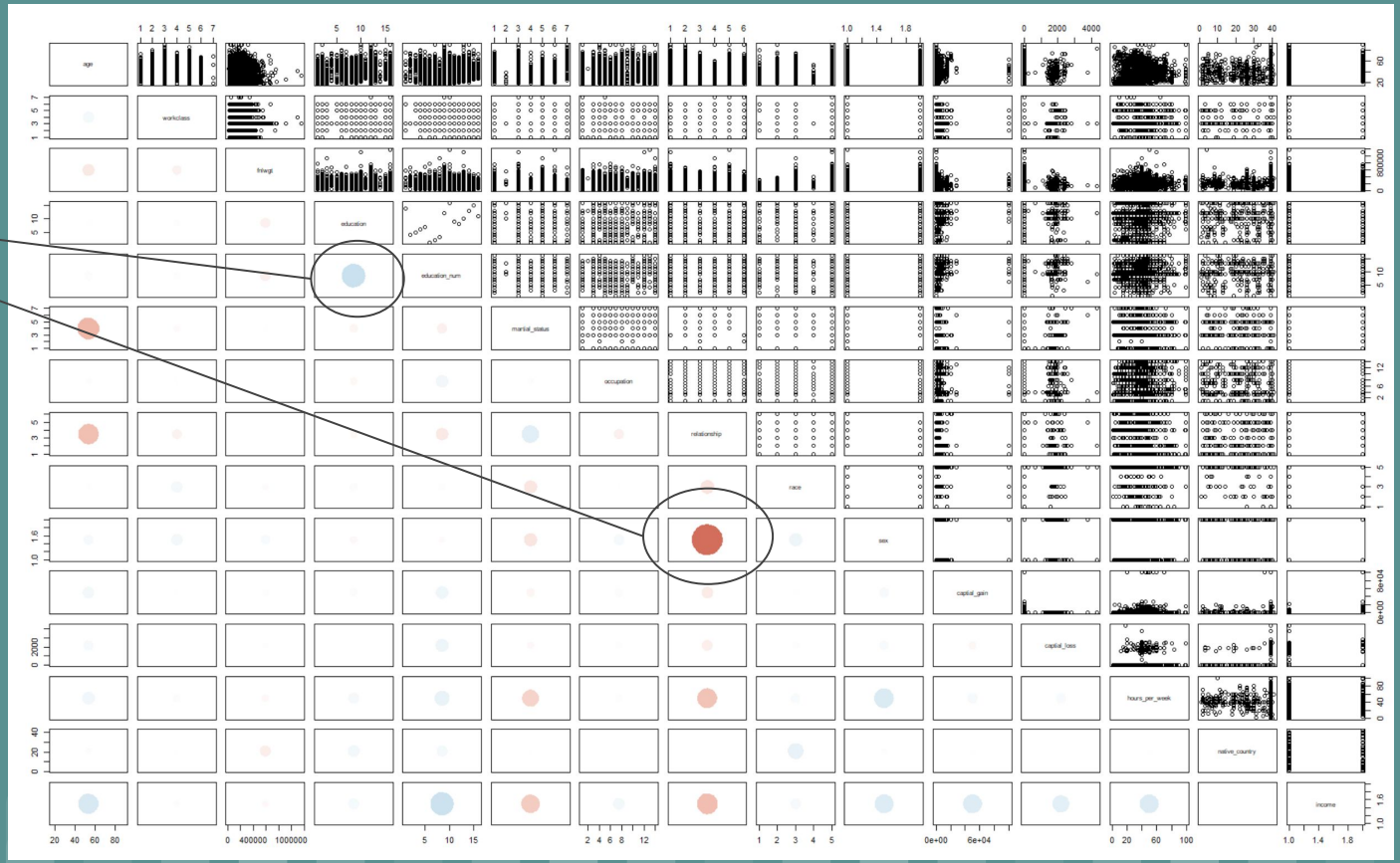
目录

数据预处理

1. 缺失值
    a. 表现为“？”
    b. 在训练集中占7.3%，测试集中7.4%
    c. 综上，缺失值数据直接删除
2. 重复值
    a. 重复数据共23条
    b. 无法确认为同一人，且量少，保留
3. 其他
    a. 无新的factor level出现在测试集中
    b. 无因错字造成的重复的factor level

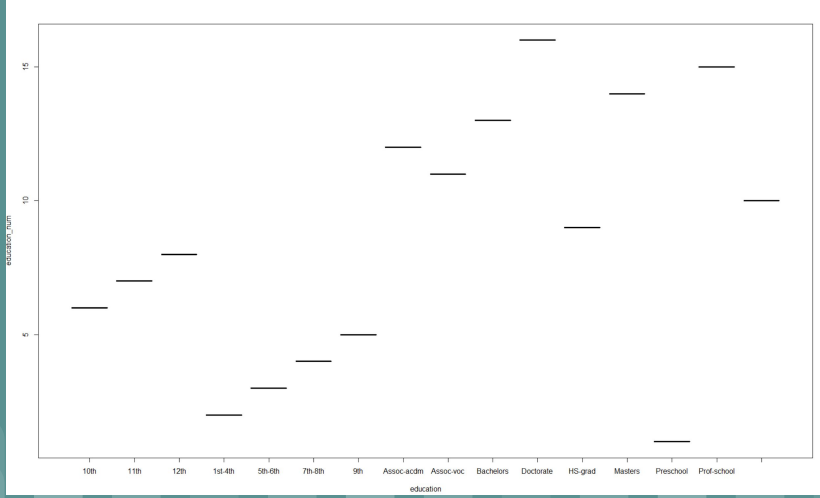
数据探索
- 特征相关

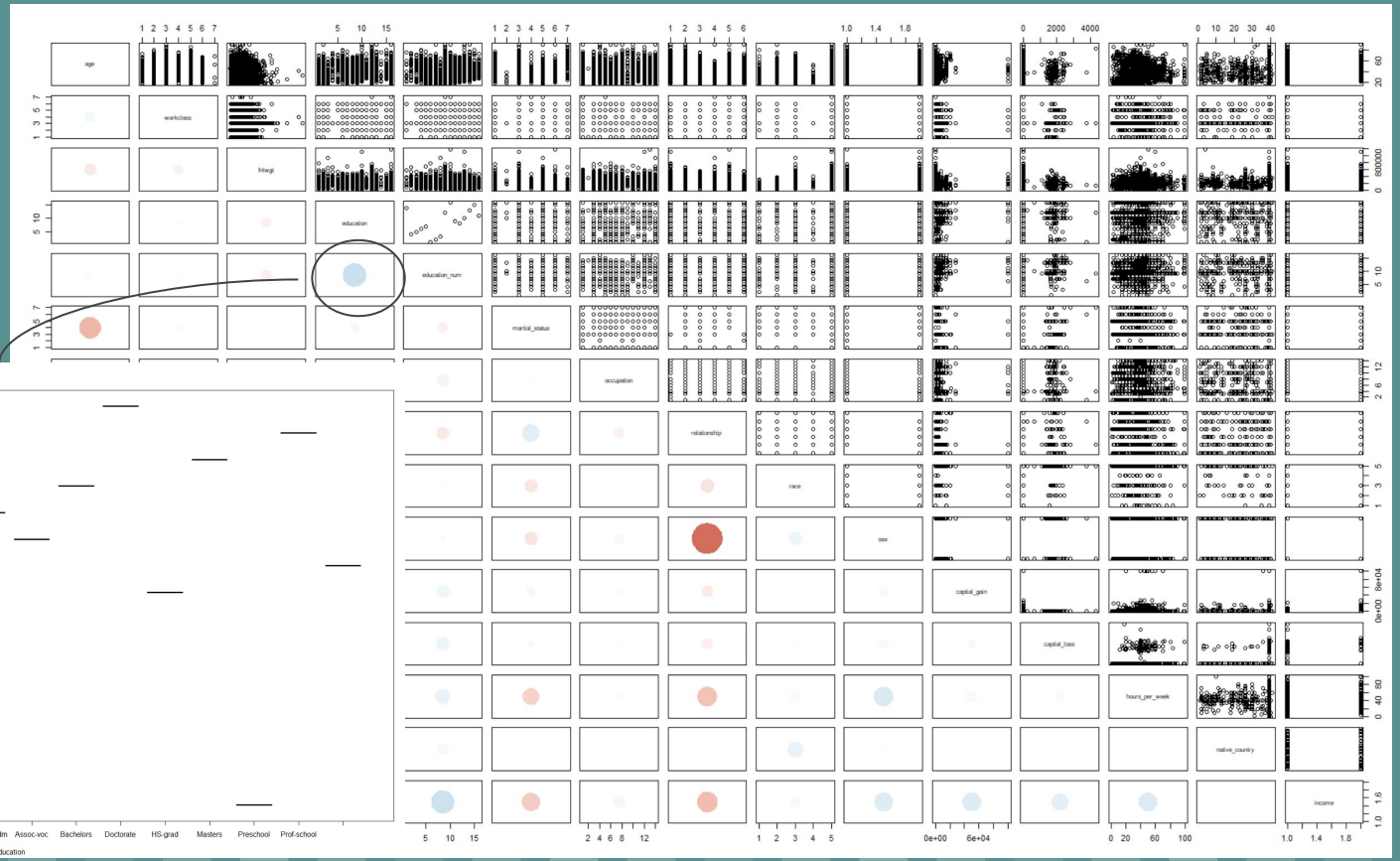
有相关性较强的特征项



数据探索
- 特征相关

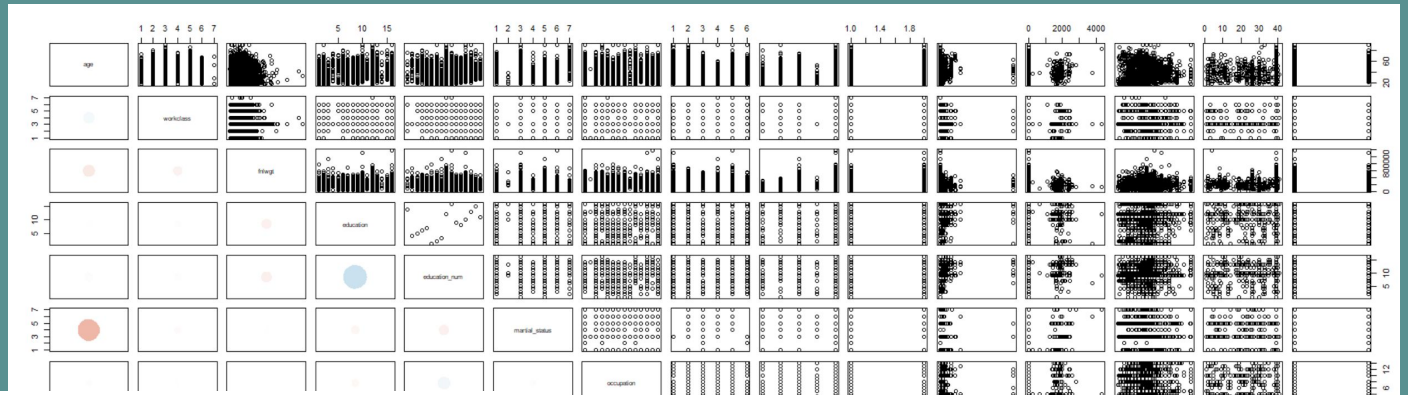
education及education_num的值俩俩对应，因此完全相同，因教育程度有叠加性，保留education_num

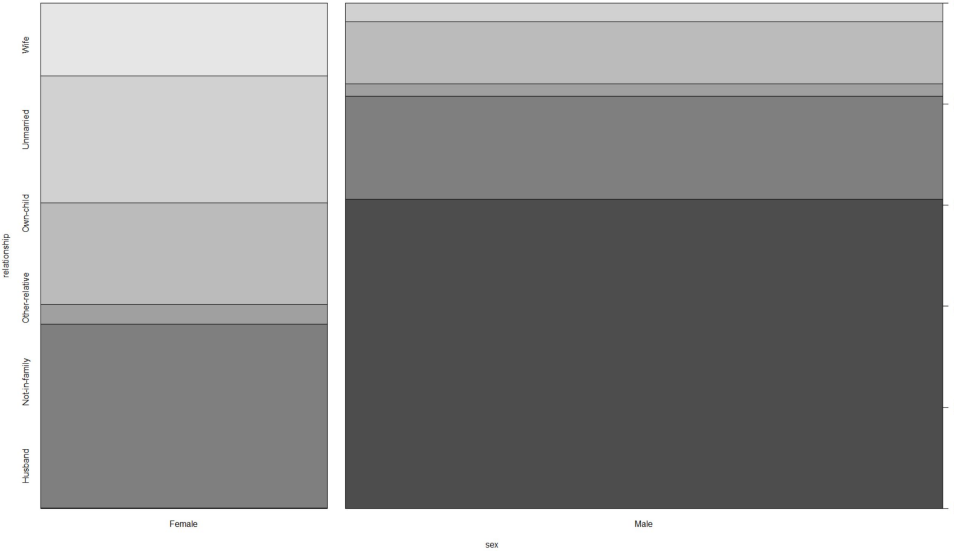




数据探索
- 特征相关

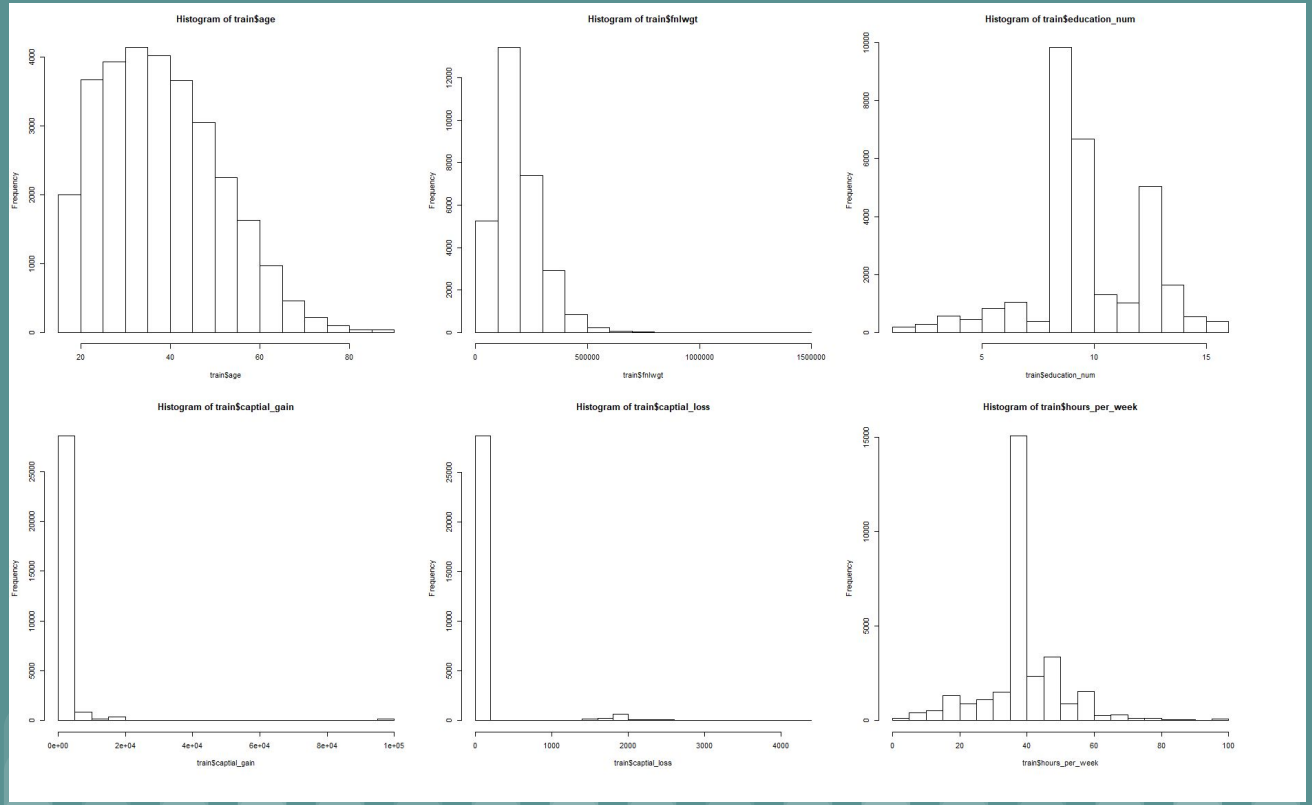
虽然相关，却不等同，
同时保留





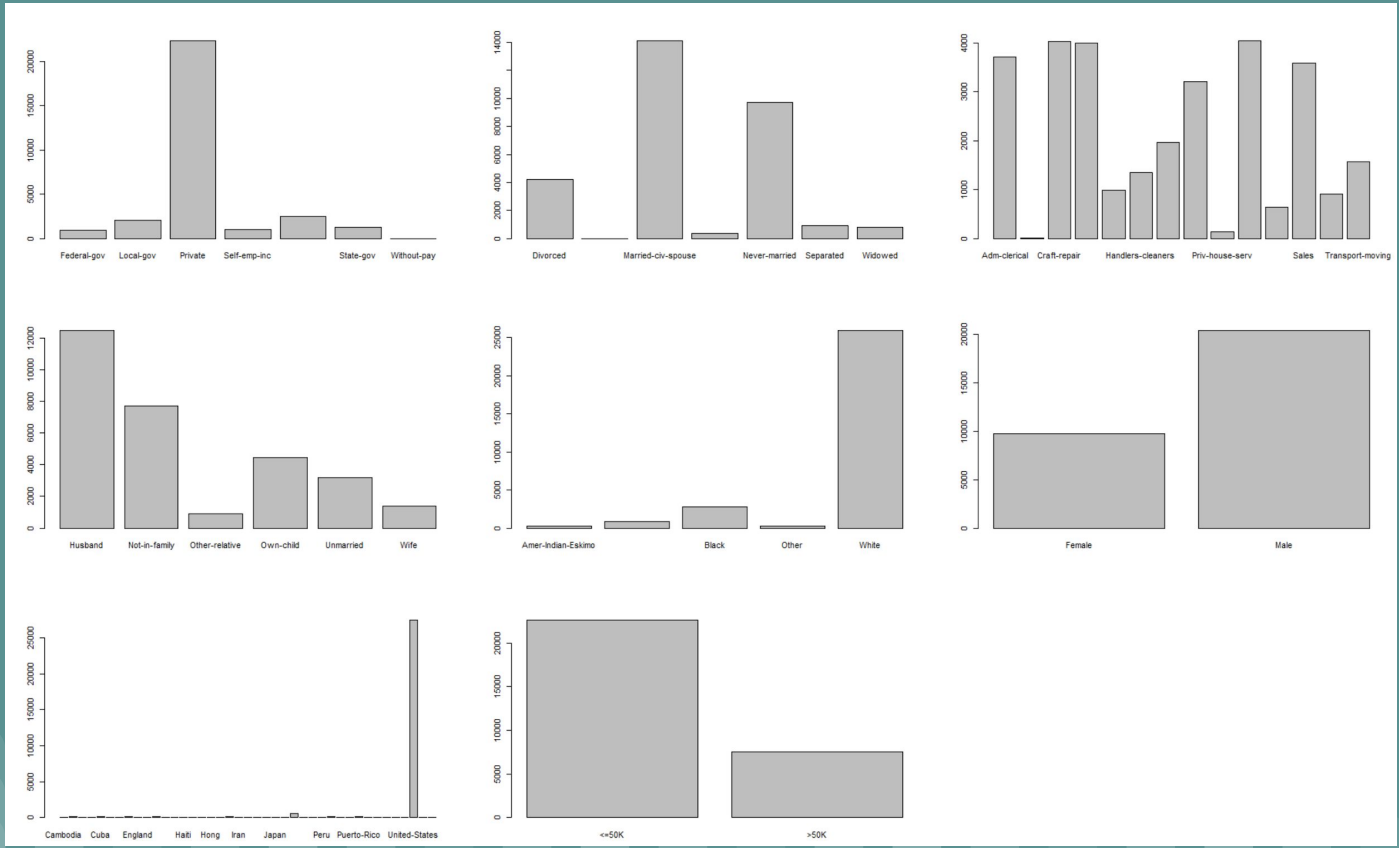
数据探索
- 特征分布

数值特征项分布:
有倾斜分布/类正态分布



数据探索
- 特征分布

类别特征项分布：
普遍分布不均



数据预挖掘

- Logistic Regression

1. 理由
    a. 二项分类
    b. 有类别特征项 - 视为dummy variable
    c. 特征项之间独立性较弱
    d. 样本量足够
    e. 最大似然促使异常的影响被压制，降低特征项倾斜性分布的影响
    f. 容易理解

数据预挖掘

- Logistic Regression

2. 模型及其分析

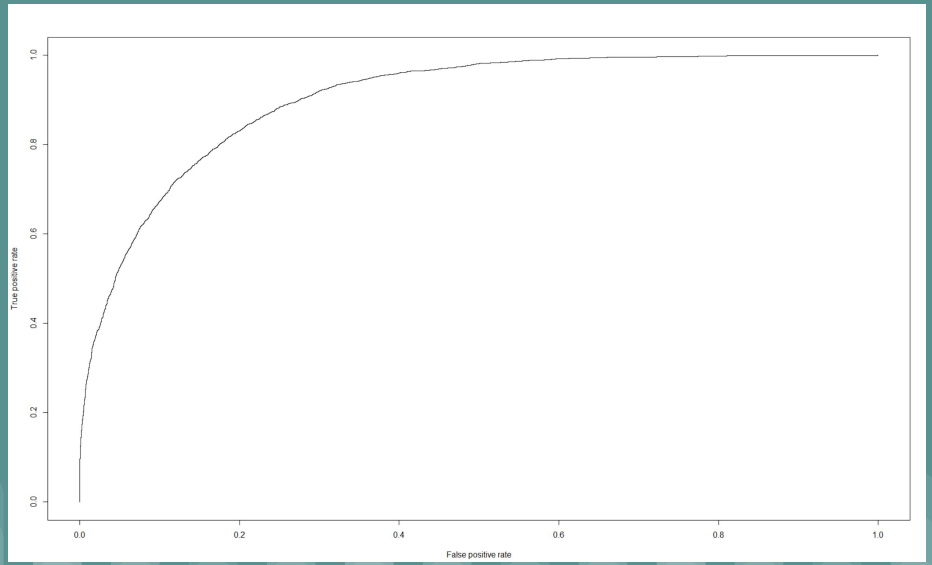
- 最显著的特征项：: age, workclass Self-emp-not-inc, education_num, occupation Exec-managerial, relationship Wife, sex male, capital_gain, capital_loss, hours_per_week
- 所有数值特征项显著
- 所有类别特征项总体显著
- 系数: 以age为例, age每增加一个单位，log odds增加0.02634
- 与Null Model显著差别
- 所有变量有对减少deviance of residuals有显著影响
- 综上，可以认为数据很好的fit这个模型

数据预挖掘

- Logistic Regression

3. 测试

准确率:0.847



AUC = .902