# 课内实验1 音乐流行度预测

58121124 张博彦

2023年4月24日

* 1. 实验设置，包括实验评估方法与性能度量
  2. 实验结果，包括模型性能，可视化结果展示
  3. 结果分析
  4. 代码附录

1. 预处理
   1. 离散属性处理

在该实验所有属性中共有3个离散属性，其中audiomode变量为哑变量无需进行处理，对剩余两个变量进行独热编码，调用函数pandas. get\_dummies进行处理

* 1. 连续属性归一化

对于所有连续属性调用MinMaxScaler().fit\_transform函数进行归一化处理

* 1. 共线性的检测与处理

检测：使用方差膨胀因子（VIF）进行多重共线性检测

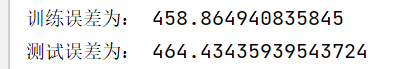
处理：对于VIF超出阈值（此处设置为5）的变量进行主成分分析（PCA）降维处理，并输出降维后的各主成分的方差值占总方差值的比例

1. 实验设置

本实验使用最基本的单次留出法，使用80%的数据集进行训练，20%的数据集进行测试并使用Python的Scikit-Learn库中的mean\_squared\_error函数分别计算训练误差（TrainError）和测试误差（TextError）。

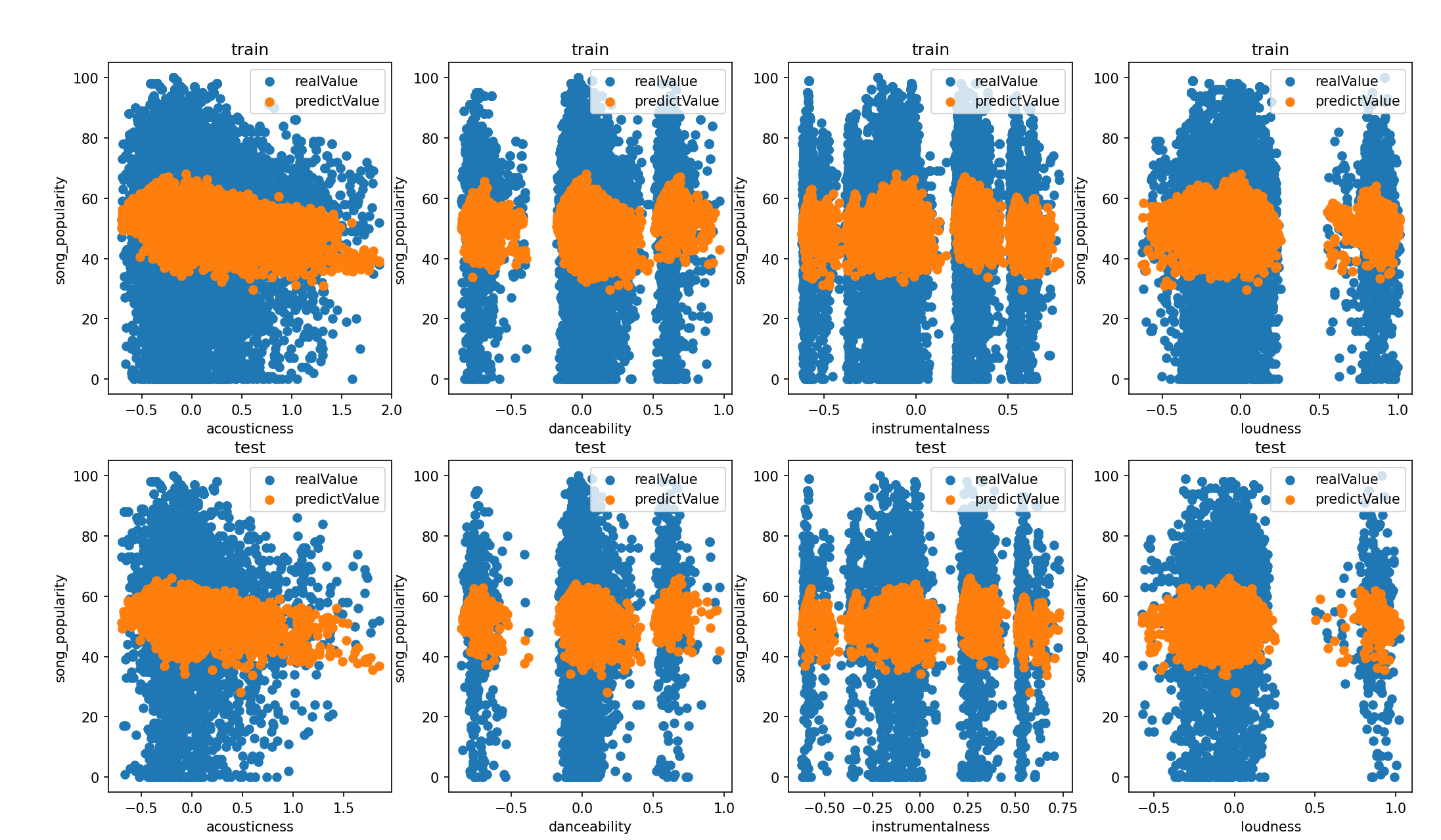
1. 实验结果

性能度量结果：



可视化结果展示：

由于自变量X维度过高，故通过计算相关系数（使用DataFrame函数）选出重要性最高的四个属性acousticness, danceability, instrumentalness, loudness分别在训练集和测试集上使用Python的Matplotlib库进行可视化



1. 结果分析

根据输出结果发现，在训练集和测试集上的MSE值都偏高，说明对于该数据集，使用线性回归模型进行拟合时误差过大，可能需要选择其他模型进行拟合

1. 代码附录

