**课内实验1 音乐流行度预测**

指导教师：刘胥影 助教：陈诗泽

学院：人工智能学院

学号：58120325

姓名：汪啸炎

1.实验环境

PyCharm Community Edition 2021.3.3

2.数据集预处理

a）离散属性连续化

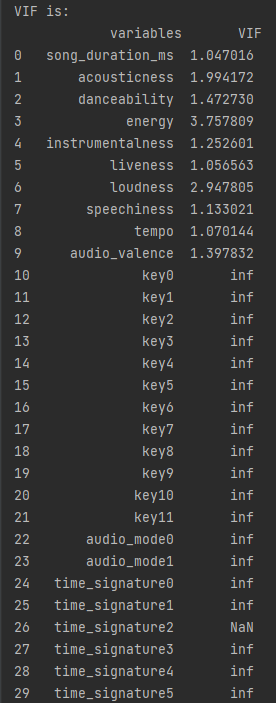
使用了独热编码，将原数据集中的key, audio\_mode, time\_signature三个属性值进行独热编码。由于它们三个分别拥有12个，2个和6个离散属性值，进行独热编码后，把key分为"key0", "key1", "key2", "key3", "key4","key5", "key6", "key7", "key8", "key9", "key10", "key11"共12个新的属性值，audio\_mode和time\_signature也进行同样处理，得到"audio\_mode0", "audio\_mode1"和"time\_signature0", "time\_signature1", "time\_signature2","time\_signature3", "time\_signature4", "time\_signature5"。这样子操作方便后续数据处理和模型训练。

b）连续属性归一化normalization

使用MinMaxScaler()，将"song\_popularity", "song\_duration\_ms", "acousticness", "danceability", "energy", "instrumentalness","liveness", "loudness", "speechiness", "tempo", "audio\_valence"这些属性进行归一化，将属性值归到[0，1]区间内，而key, audio\_mode, time\_signature三个属性由于进行了独热编码，处理后得到的属性值非0即1，不必再进行归一化处理。

c）共线性的检测与处理

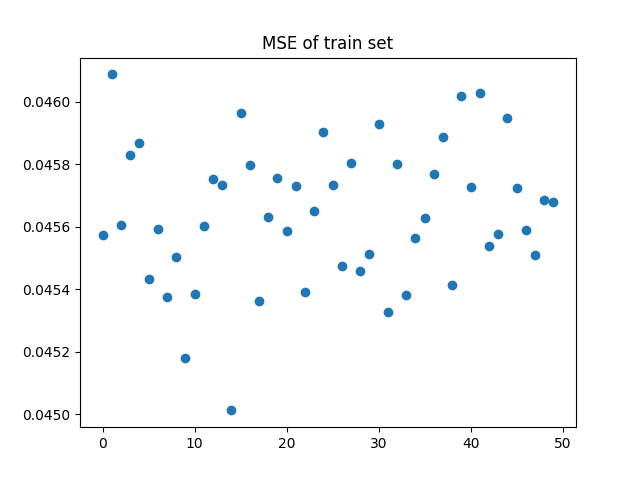
先计算了VIF，得到的结果如下图所示，根据文档提示，不必进行共线性处理。

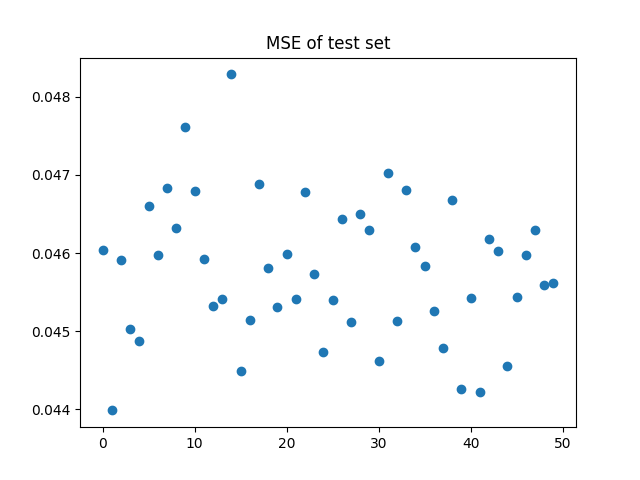


2.实验设置

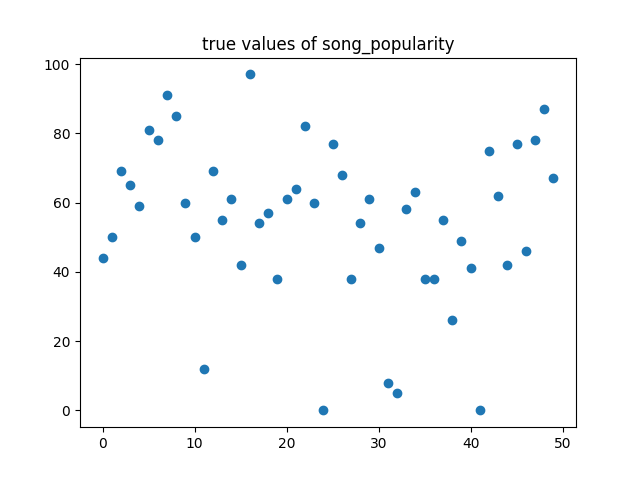
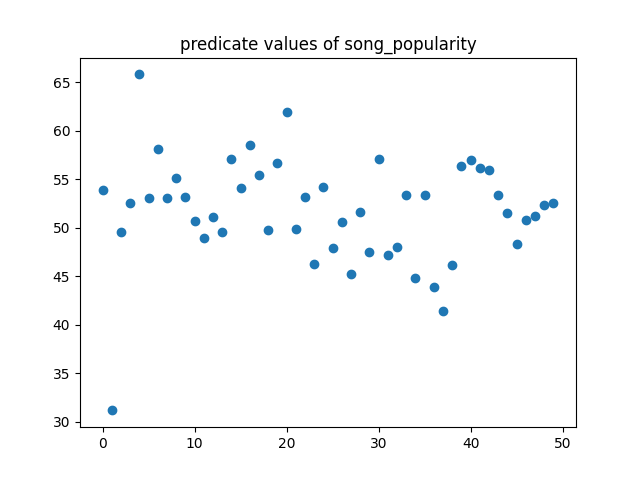
首先用train\_test\_split()函数拆分数据集为训练集和测试集，模型训练结束后，再利用MSE计算测试集误差和训练集误差。

3.实验结果

通过随机划分训练集和测试集，进行了50次实验，并分别计算出每一次实验的误差，最后将它们可视化，得到了如下结果：



其中一次实验，从中取得了前50个数据的song\_popularity真实值和预测值，并将它们可视化，得到了如下结果：



4.结果分析

从实验结果可以看出，训练模型所得到结果的均方误差（MSE）都在0.05以内，表现较好。但是，从song\_popularity的真实值和预测值来看，预测值集中在[40,60]区间内，对最值的表现不佳。可能的原因是，在对离散属性进行处理时，产生的稀疏值过多，导致训练出来的的模型考虑因素过多，每个因素权值下降，对song\_popularity的评分预测趋于保守。

5.代码附录见文件夹code

附：虽然经过计算VIF值，得到了不必进行共线性处理的结论，但在代码中我还是试着进行了PCA降维操作，不过并没有通过降维改变训练数据。