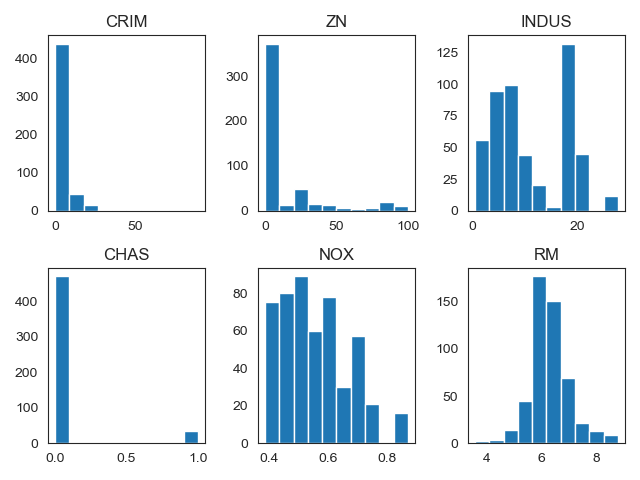
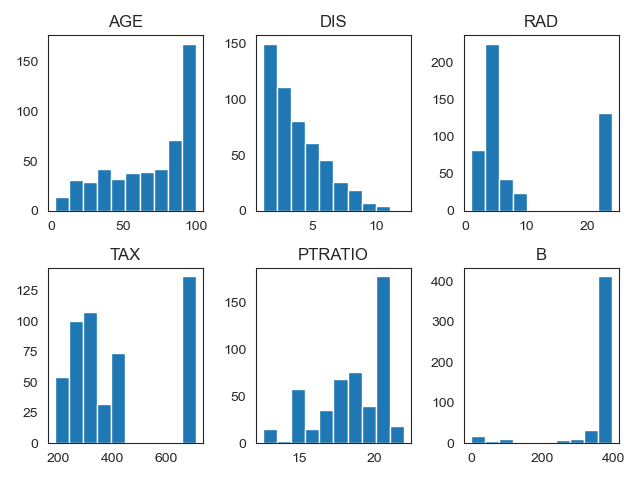
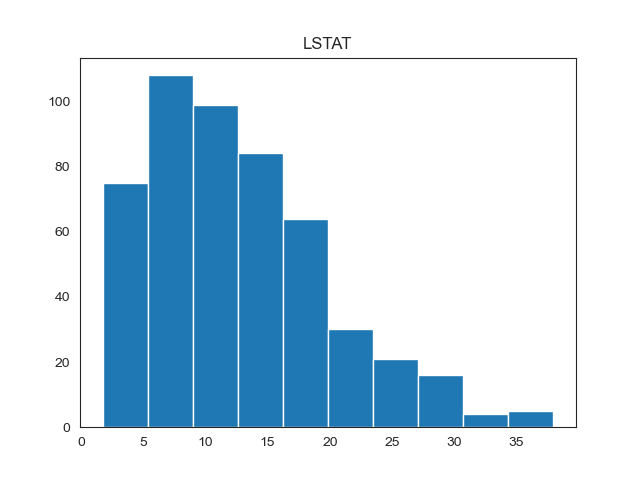
1. 特征分布可视化

对13个特征的分布情况进行可视化：  


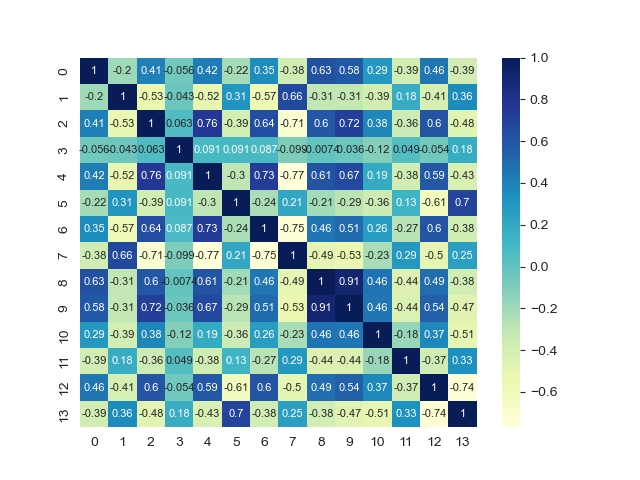




可以看到，每个特征的数据分布均不属于正态分布，同时，每个特征之间的取值范围存在差异，需要对数据进行归一化操作，使得特征范围一直，且服从正态分布，便于进行模型训练。

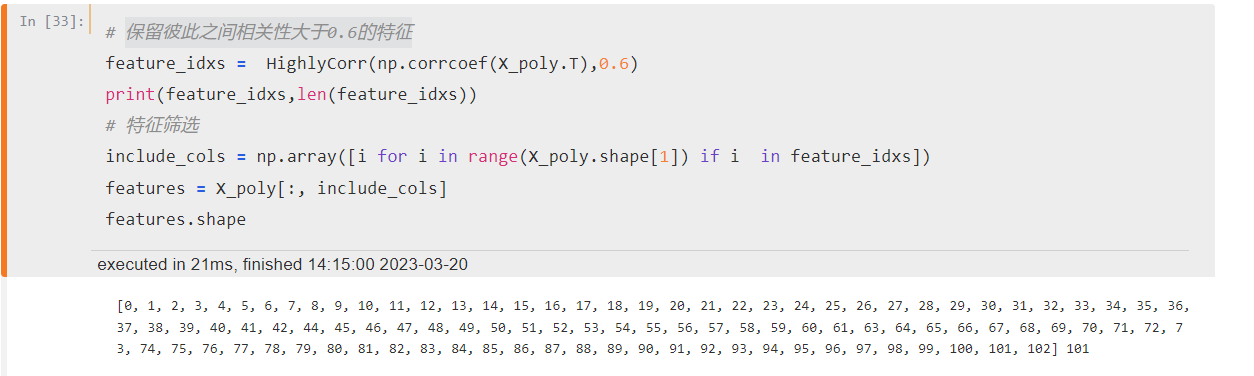
1. 特征相关性可视化

如图，可以看到有些特征与其他特征和预测变量的相关性较弱，如feature3，推测可能是由于该特征的0取值过多，可以考虑剔除该特征。



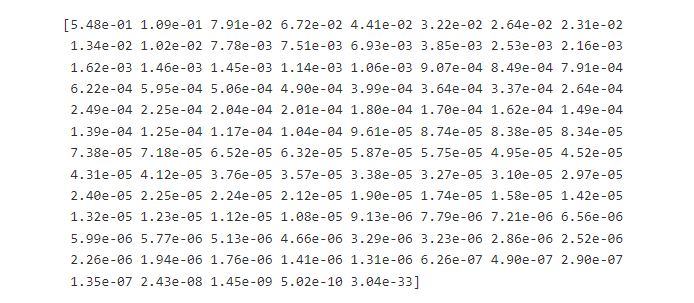
1. 特征组合

使用PolynomialFeatures()，设置degree=2，生成更多特征，并保留保留彼此之间相关性大于0.6的特征，共保留101个特征。



然而过多的特征会导致模型出现过拟合，因此采用PCA主成分分析进行降维：

首先，查看101个特征在PCA后各自解释方差的大小



然后，选取前30个特征，后续的特征对方差的解释过小，几乎可以忽略，前30个特征足以描述数据的全部特征。

1. 模型训练结果

模型实现见main.py文件，这里没有使用上述预处理生成的数据进行训练，而是对原数据进行归一化后直接划分数据进行训练和验证。

模型结果可视化如下：在迭代5000轮后，train loss 收敛至4.76

test loss收敛至:8.29

