# 工程数据分析文稿

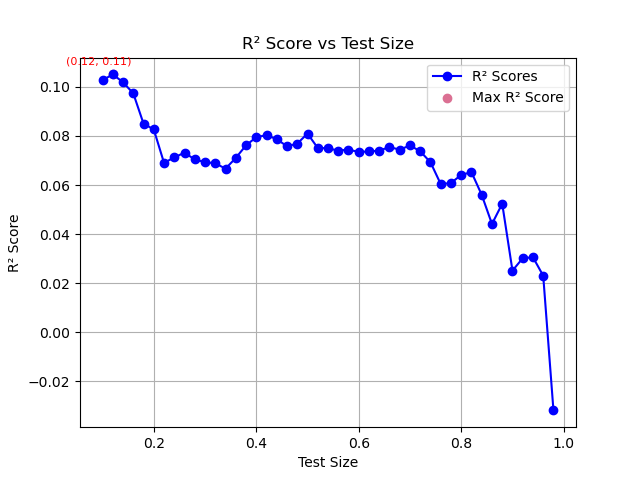
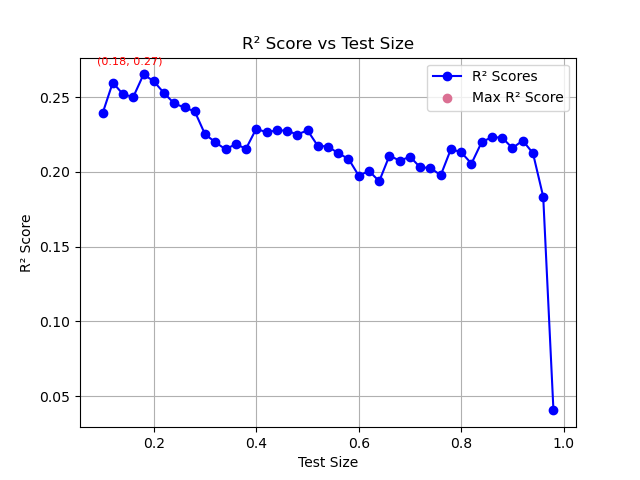
## No.1多元线性回归模型

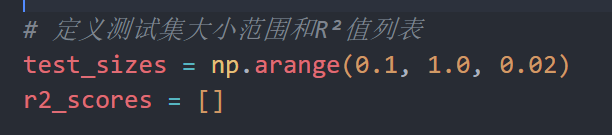
### 直接预测（数据未处理）：

多元线性回归模型是指含有多个解释变量的线性回归模型，用于解释被解释的变量与其他多个变量解释变量之间的线性关系。其数学模型为：

我们使用sklearn.linear\_model.LinearRegression类创建了一个线性回归模型实例。在这个例子中，我们使用了5个自变量（Thickness', 'Milling depth', 'Perception', 'Freeze Index','Structual number','Age'）来预测因变量（'TN'/’A’列）。多元线性回归模型允许我们使用多个自变量来预测一个因变量

##### 测试集和训练集的分配





以上分别是以0.02为步长最终生成的测试机占比和**R2**之间关系的点线图，由此可得，当测试集占比为0.18（TN）和0.12（A）时两者的**R2**分别处于最大，这时我们按照此标准分别对两组数据训练

得到，两者对应的R2和MSE分别为

TN

Mean Squared Error: 444.2745806853906

R² Score: 0.26528343632368534

A

Mean Squared Error: 2927.720222504276

R² Score: 0.10508774088843609

##### 指标解释和模型评估

根据 R-Squared 的取值，来判断模型的好坏，其取值范围为[0,1]：

如果结果是 0，说明模型拟合效果很差；

如果结果是 1，说明模型无错误。

对TN和A值的预测R2都比较低，所以拟合效果并不理想

## N0.2参考鲍鱼年龄预测的机器学习回归模型

##### 1数据的归一化处理

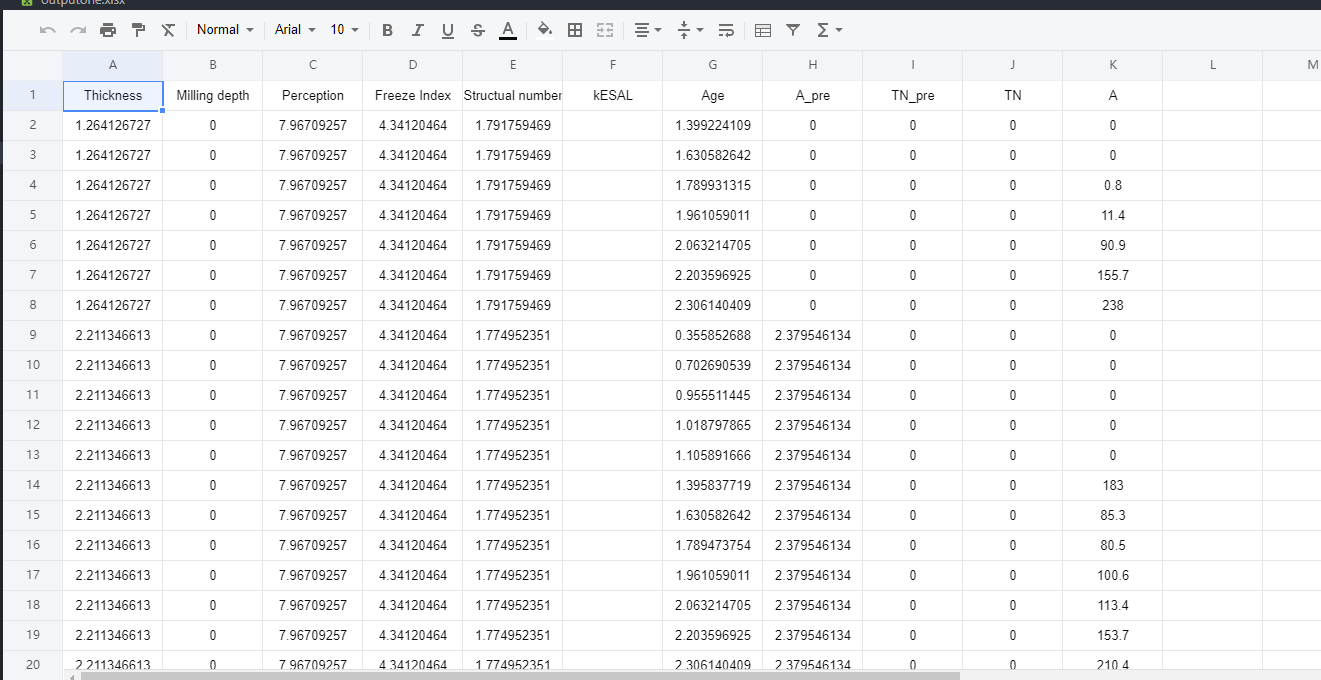
建立回归模型之前，不同评价指标往往具有不同的量纲和量纲单位，这样的情况会影响到数据分析的结果，对数据进行归一化处理使得各指标处于同一数量级从而提升训练和与预测的性能。

###### 判断与选择

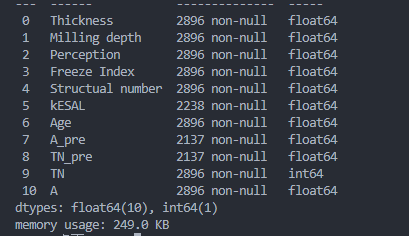
通过计算数据集中每个特征的最大值与最小值的比值来判断是否需要归一化并且计算数据的偏度来选择合适的归一化方法。如果数据偏度较大（绝对值平均大于2），建议使用对数变换来减小偏度。否则，建议使用MinMaxScaler进行归一化处理

最终运行结果使用Log transformation归一化处理

最终得到的outputone.xls文件就是对自变量归一化后的文件

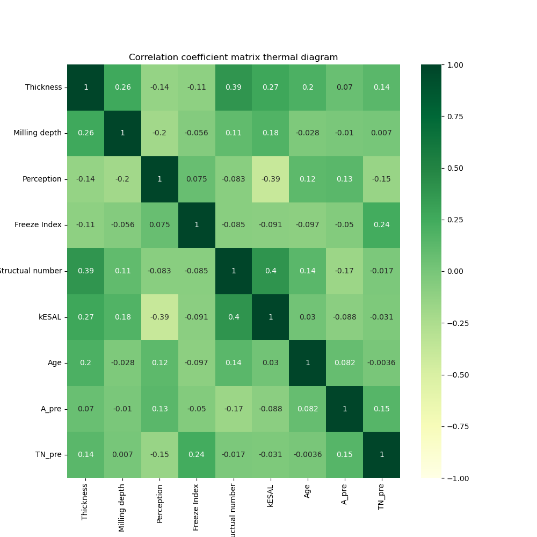
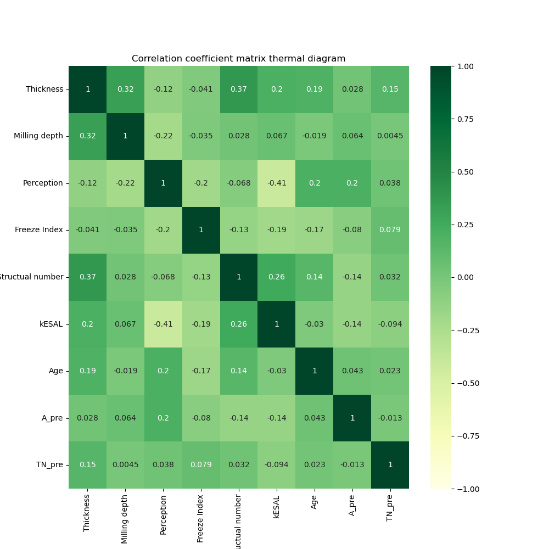


接下来对数据进行特征值和缺失值检测



TN和A作为因变量/目标变量/因此存在9个特征值

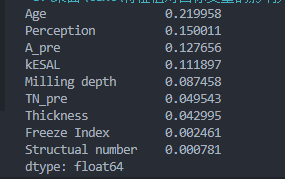
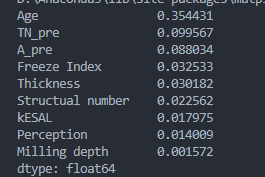
特征值矩阵热力图



由此可见，特征值之间的相关系数都在0.5以下，说明特征值之间的线性关系较弱

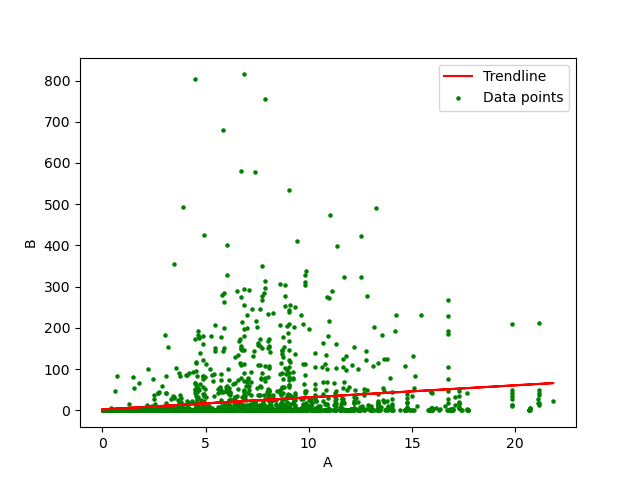
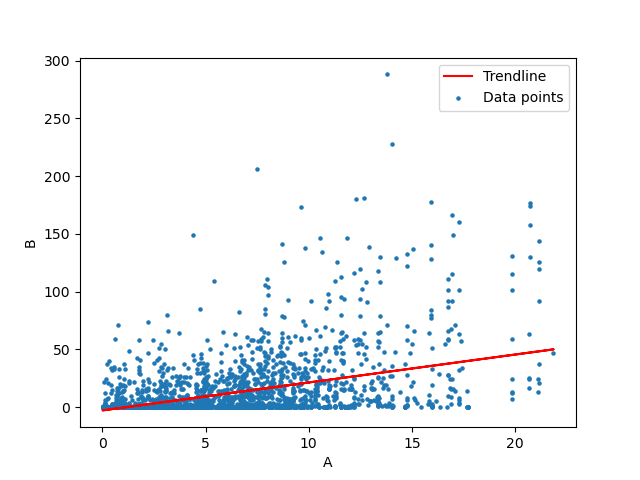
评判不同的特征值对结果的影响因素大小

下图分别是针对目标变量为TN和A时，特征值的影响指数大小

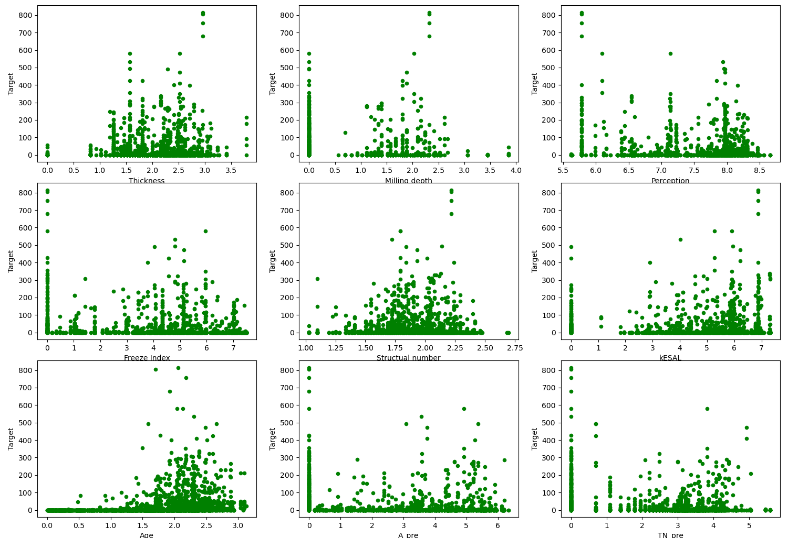
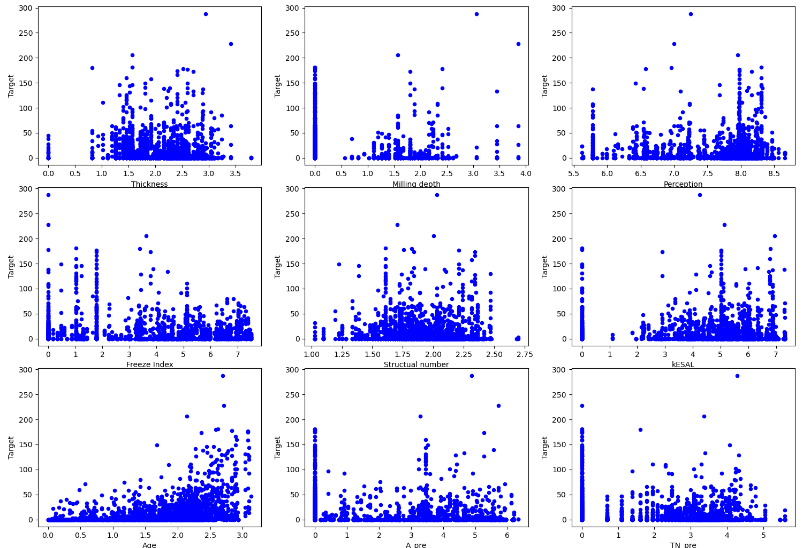


由此可以得出Age特征值对目标变量的影响最大

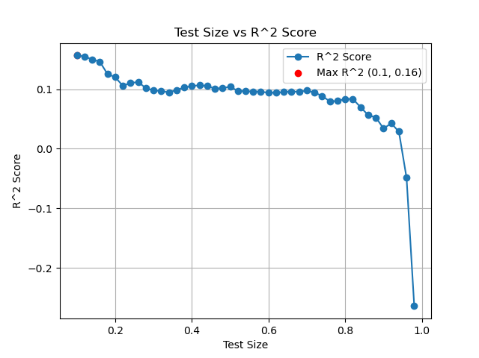
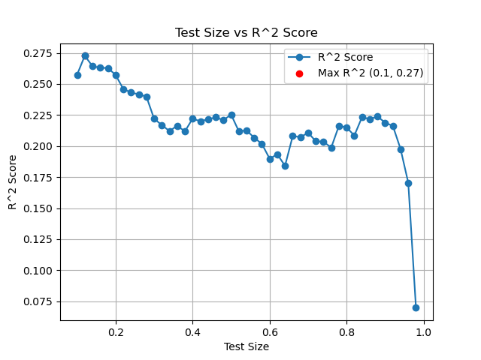
Age特征与TN以及A的散点图与趋势线



下图是九个特征值各自对目标变量TN和A的散点图



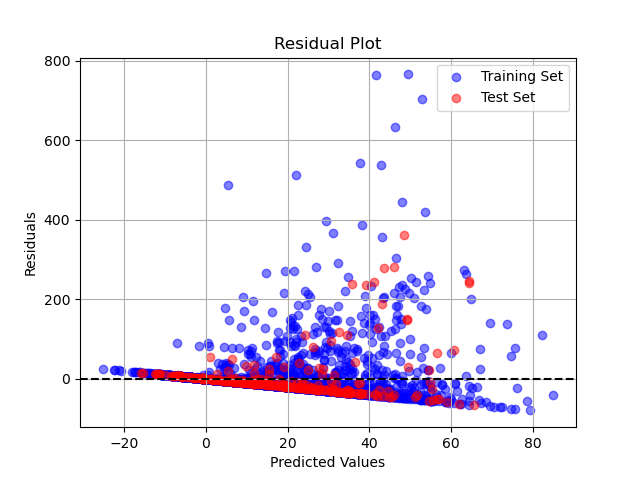
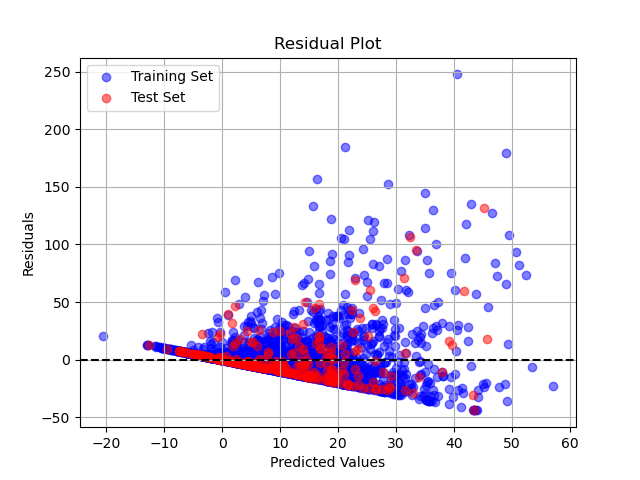
Lasso回归预测：



残差图：

残差图是一种用来诊断回归模型效果的图。在残差图中，如果点随机分布在0附近，则说明回归效果较好。如果在残差图中发现了某种结构,则说明回归效果不佳。

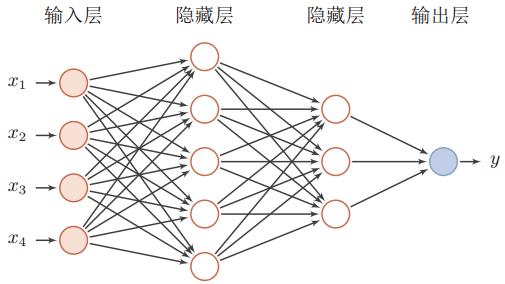
以下分别是

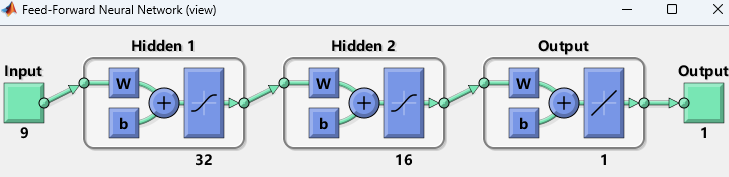


观察残差图的分布情况，理想情况下，残差应当随机分布在Y=0附近，没有明显的模式。如果残差图中存在某种模式或趋势，那么可能说明模型未能捕捉到某些重要特征或存在异方差问题。原因可能是未能足够对数据进行预处理。

困了，明天再收一下尾

深度学习——多层前馈神经网络的预测分析

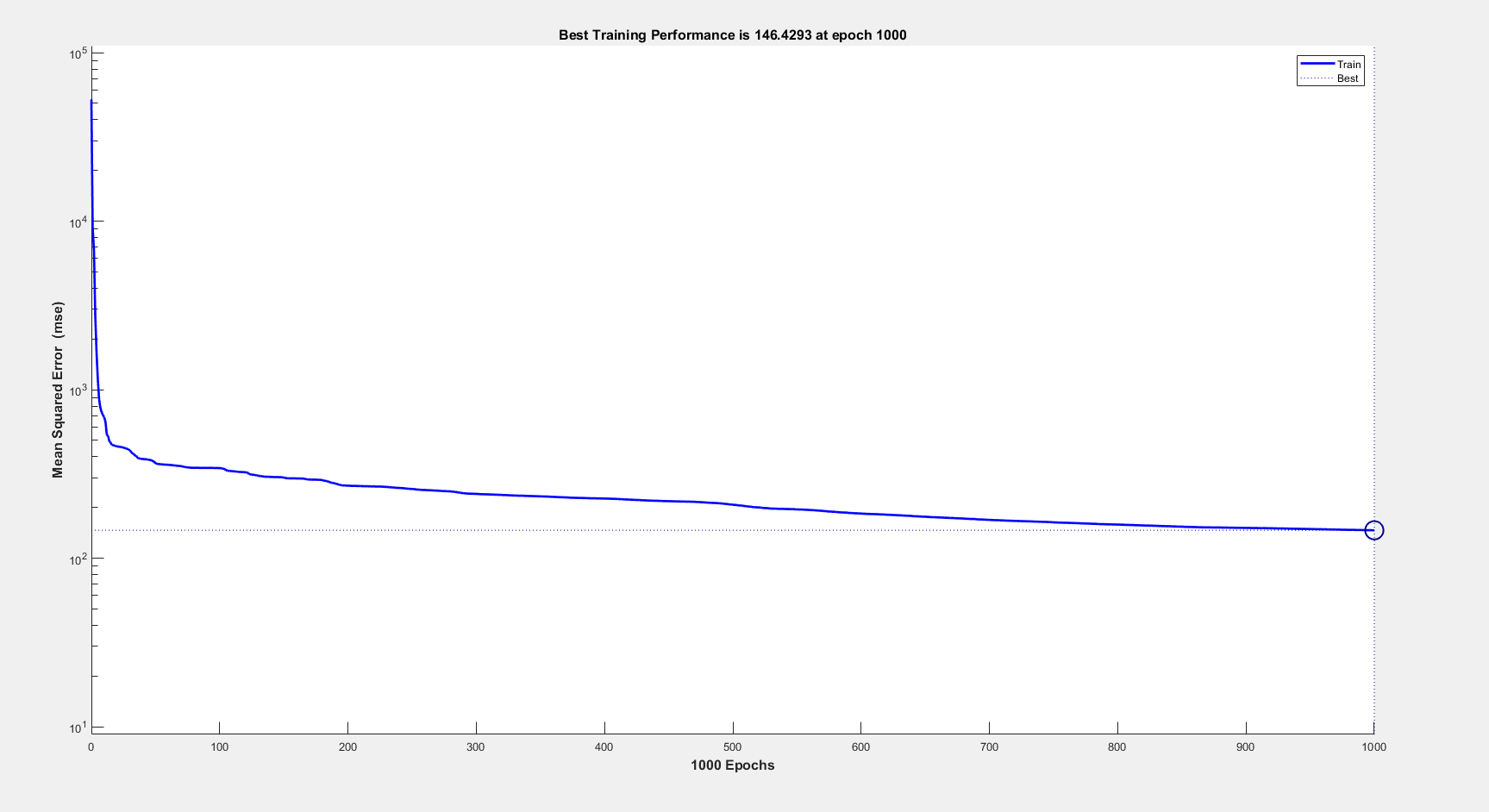
多层前



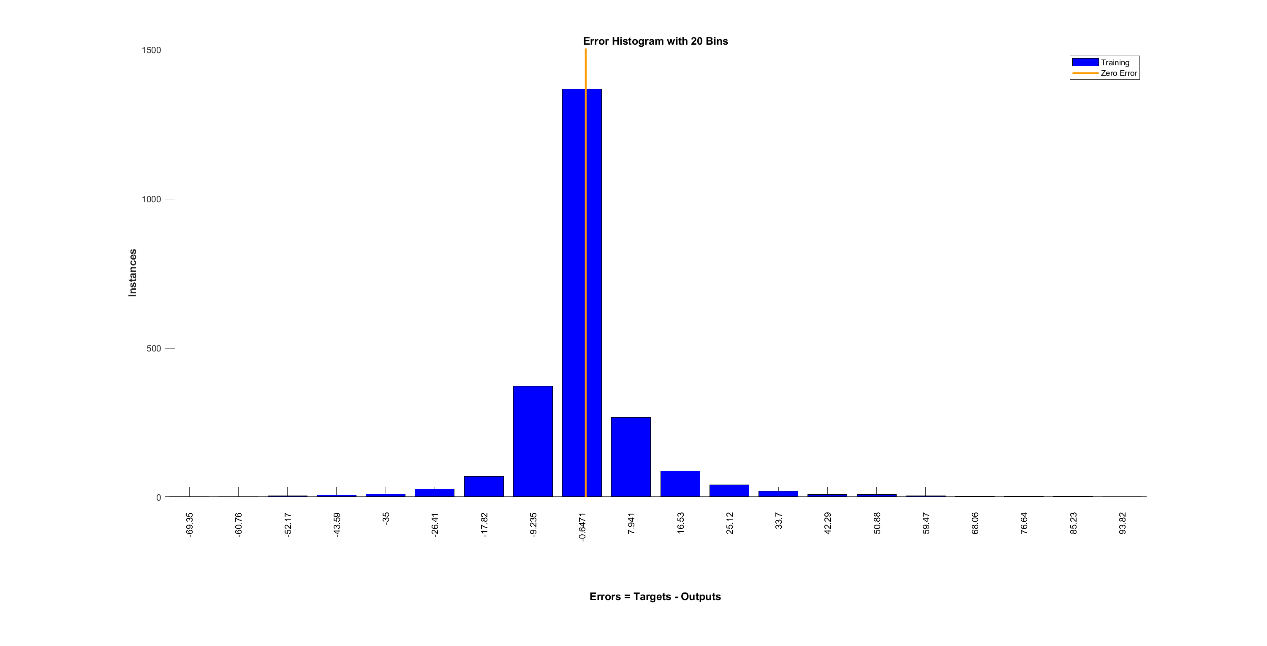
这种神经网络结构包括一个输入层、若干隐藏层和一个输出层。输入层接收特征数据，隐藏层用于处理和学习数据中的复杂关系，输出层提供最终的预测值。

在这个示例中，我们使用了两个隐藏层，它们的神经元数量分别为32和16。隐藏层使用双曲正切激活函数（tansig），输出层使用线性激活函数（purelin）。使用均方误差（MSE）作为性能度量，采用训练策略为Scaled Conjugate Gradient（trainscg）。

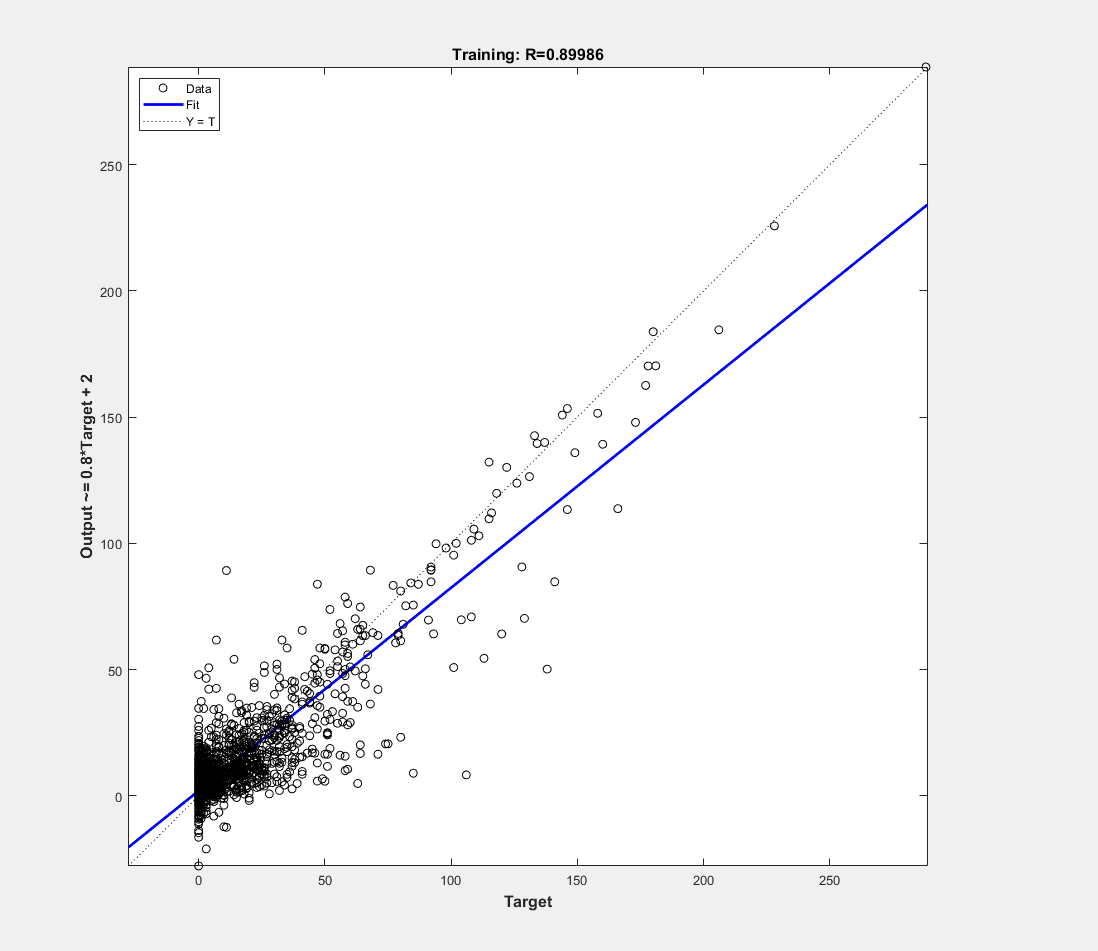
训练与评估



MSE减小并最终趋于10的平方。这表明神经网络在训练过程中逐渐改善了预测性能。



横坐标表示误差的范围，纵坐标表示落在该误差范围内的样本数量。每个柱子代表一个特定误差范围内的样本数量。大部分误差应该集中在0附近，表示预测值与真实值非常接近。



散点图是·真实值和预测值之间关系的散点图

将神经网络的预测输出值与真实目标值进行比较。横坐标表示真实目标值，纵坐标表示神经网络的预测值。

图中R = 0.89986，说明真实值和预测值之间存在很强的相关关系，模型的回归状况优良

R-squared (R2): 0.50678，说明预测结果优良