# 构建机器学习模型及训练

本章将使用第五章构建的特征提取器对样本数据构建数据集，并对数据集样本进行分析，洞见特征数据间的数学联系及统计学关系。

## 数据集分析

### 数据集构建

由于机器学习本质上是使用大量的样本数据，通过计算机实现的数学模型，来探索并计算输入特征之间的数学关系，从而实现预测，所以需要使用大量的样本数据来构建数据集，所以我们使用300个Verilog HDL设计及其在Xilinx Vivado综合工具上综合后输出的资源利用率结果来作为基本样本。

通过使用第五章构建的特征提取器来对样本数据进行提取，并输出至CSV，我们可以得到一个基本数据集**，如图所示：**

### 数据集分析

#### 数据可视化

我们将数据进行可视化，将可以便于我们对数据进行进一步的分析，使用直方图来展示数据：**如图：**

#### 寻找相关性

数据(data)是事实或观察的结果，是对客观事物的逻辑归纳，是用于表示客观事物的未经加工的的原始素材。数据相关性是指数据之间存在某种关系，如正相关，负相关。所以我们需要对数据进行相关性分析，指标为标准相关系数也称为皮尔逊系数r。

与LUT资源的相关性如下：

与FF资源的相关性如下：

与IO资源的相关性如下：

与BUFG资源的相关性如下：

## 模型选择

在完成数据分析，特征分析，特征工程后，我们需要选择合适的机器学习模型来进行预测，首先我们不知道选择哪种模型的时候，先对各种回归模型进行初步尝试，指标采用R2和RMSE进行评估；

### 初步模型选择

我们分别使用线性回归、决策树、SVM向量机、随机森林、深度神经网络来对数据集进行初步的训练，从而选择出两种有前途的特种进行进一步的训练和微调从而实现我们的预测目标。

首先我们对数据集进行划分，我们随机抽样20%的数据作为验证集，剩余80%作为训练集，为保证训练及验证的全面覆盖，我们在验证时将使用5折交叉验证来对我们的模型进行预测验证。

**以下结果均采用标准参数进行训练：如图所示**

通过上图所示，我们可以发现在线性回归、决策树的性能较差模型预测表现一遍，但是随机森林、SVM向量机、深度神经网络这三种模型在标准参数下存在较好的的预测性能，所以本文下面将着重以随机森林及深度神经网络来进行进一步的预测模型训练。

## 随机森林

## 深度学习

## 本章小结

**本章主要叙述了对多种机器学习模型在标准参数下对使用文本方法所构建的数据集的预测性能表现及对随机森林、深度神经网络进行进一步预测、参数微调等方法及结果的叙述，最终我们得到了对RTL设计在FPGA片上资源利用率预测的较为满意的结果的模型。**