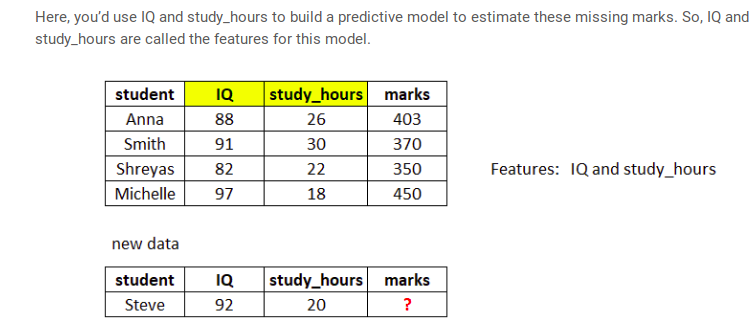
特征：

在机器学习的环境中，一个特征可以被描述为解释现象发生的一个要素或一组要素。当这些要素被转换成某种可测量的形式时，它们被称为特征。

例如，假设你有一份学生名单。这个列表包含每个学生的名字，他们学习的小时数，他们的智商，以及他们在以前考试中的总成绩。现在你会得到一个新学生的信息——他/她学习的小时数和智商，但是他/她的分数不见了。你必须估计他/她的可能分数。

在这里，你可以用智商和学习时间来建立一个预测模型来估计这些缺失的分数。因此，智商和学习时间被称为这个模型的特征



特征工程

现在要创建一个新特性，我们可以使用项目权重和项目价格。因此，让我们创建一个名为“每磅价格”的特性。它只是商品的价格除以商品的重量。这个过程称为特征工程。

这只是从现有特性中创建新特性的一个简单例子，但是在实践中，当我们拥有大量特性时，特性工程会变得非常复杂和麻烦。

让我们举另一个例子。在流行的泰坦尼克号数据集中，有乘客姓名功能，下面是数据集中的一些姓名:

为什么需要特征工程？

预测模型的性能在很大程度上取决于用于训练该模型的数据集中特征的质量。如果您能够创建新的特性来帮助向模型提供更多关于目标变量的信息，那么它的性能将会提高。因此，当我们的数据集中没有足够高质量的要素时，我们必须依靠要素工程。

智能特性工程有助于确保在排行榜前5%的位置。创建的一些功能如下:

小时箱:宁滨在决策树的帮助下创建了一个新的小时功能

温度箱:类似地，温度变量的箱特征

年份箱:创建8个季度箱，为期2年

日类型:日分为“工作日”、“周末”或“假日”

创建这样的功能绝非易事——它需要大量头脑风暴和广泛的数据探索。不是每个人都擅长特征工程，因为它不是你可以通过读书或看视频来学习的东西。这就是为什么特征工程也被称为一门艺术。如果你擅长它，那么你在竞争中有很大优势。

Featuretools工具：

Featuretools是一个用于执行自动化特征工程的开源库。这是一个很好的工具，旨在快速推进特征生成过程，从而给予更多时间关注机器学习模型构建的其他方面。换句话说，它让你的数据“机器学习准备就绪”。

三要素：

**Entityset**

**Deep Feature Synthesis (DFS)**

**Feature primitives**

实体可以被视为Pandas DataFrame的表示。

多个实体的集合称为Entityset。

DFS:

深度特征合成（DFS）与深度学习无关。

DFS实际上是一种特征工程方法，是Featuretools的支柱。 它支持从单个数据帧和多个数据帧创建新功能。

**Feature primitives：特征基元**

DFS通过将**Feature primitives**应用于EntitySet中的实体关系来创建特征。

这些原语是手动生成特征的常用方法。 例如，原语“mean”将在聚合级别找到变量的均值。

特征选择工具feature-selector，feature-selector是由Feature Labs的一名数据科学家williamkoehrsen写的特征选择库。feature-selector主要对以下类型的特征进行选择：

具有高missing-values百分比的特征

具有高相关性的特征

对模型预测结果无贡献的特征（即zero importance）

对模型预测结果只有很小贡献的特征（即low importance）

具有单个值的特征（即数据集中该特征取值的集合只有一个元素）

特征提取：

<https://www.kaggle.com/pmarcelino/data-analysis-and-feature-extraction-with-python>

https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/