Different Types of Data Science Problems

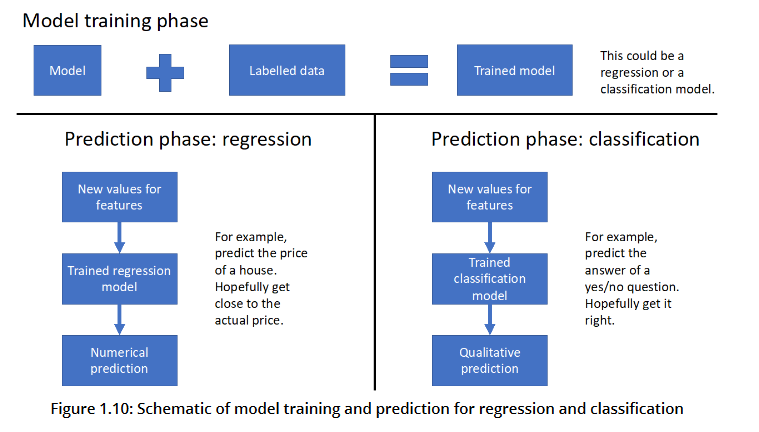
If you have a dataset of 1,000 houses including the values of these features and the prices of the houses, you can say you have 1,000 samples of labeled data, where the labels are the known values of the response variable: the prices of different houses. Most commonly, the tabular data structure is organized so that different rows are different samples, while features and the response occupy different columns, along with other metadata such as sample IDs,

1. **Regression Problem**

Once you have trained a model to learn the relationship between the features and response using your labeled data, you can then use it to make predictions for houses where you don't know the price, based on the information contained in the features. The goal of predictive modeling in this case is to be able to make a prediction that is close to the true value of the house. Since we are predicting a numerical value on a continuous scale, this is called a regression problem

1. **Classification Problem**

On the other hand, if we were trying to make a qualitative prediction about the house, to answer a yes or no question such as "will this house go on sale within the next five years?" or "will the owner default on the mortgage?", we would be solving what is known as a classification problem. Here, we would hope to answer the yes or no question correctly. The following figure is a schematic illustrating how model training works, and what the outcomes of regression or classification models might be:



Classification and regression tasks are called supervised learning, which is a class of problems that relies on labeled data. These problems can be thought of a needing "supervision" by the known values of the target variable. By contrast, there is also unsupervised learning, which relates to more open-ended questions of trying to find some sort of structure in a dataset that does not necessarily have labels. Taking a broader view, any kind of applied math problem, including fields as varied as optimization, statistical inference, and time series modeling, may potentially be considered an appropriate responsibility for a data scientist.