Improving Model Performance

To improve the performance of the model, a few of the techniques to consider are

1. Systematic feature engineering
2. Using ensemble learning methods (we'll discuss more on this in a later chapter)
3. Hyper-parameter tuning of the algorithm

Feature Engineering

In model building, a significant portion of time is spent on feature engineering. Feature engineering is the practice of systematically going through each feature in the dataset and investigating its relevance to the targets.

Through feature engineering, we can cleverly introduce new features by combining one or more existing features, and this can impact the prediction accuracy of the model. Feature engineering can sometimes be the difference between a decent learning model and a competition-winning model.

Ensemble Methods

Ensemble methods combine the output of weaker models to produce a better performing model. Two major classes of ensemble learning algorithms are

- Boosting
- Bagging

In practice, ensemble methods such as Random forests are known to do very well in various machine learning problems and are the algorithms of choice for machine learning competitions.

Hyper-parameter Tuning

When modeling with a learning algorithm, we can adjust certain configurations of the algorithm. These configurations are called hyper-parameters. Hyper-parameters are tuned to get the best settings of the algorithms that will optimize the performance of the model. One strategy is to use a grid search to adjust the hyper-parameters when fine-tuning the model.

Unsupervised Learning

In unsupervised learning, the goal is to build a model that captures the underlying distribution of the dataset. The dataset has no given targets for the input features (see Figure 14-17). So, it is not possible to learn a function that maps a relationship between the input features and the targets as we do in supervised learning.

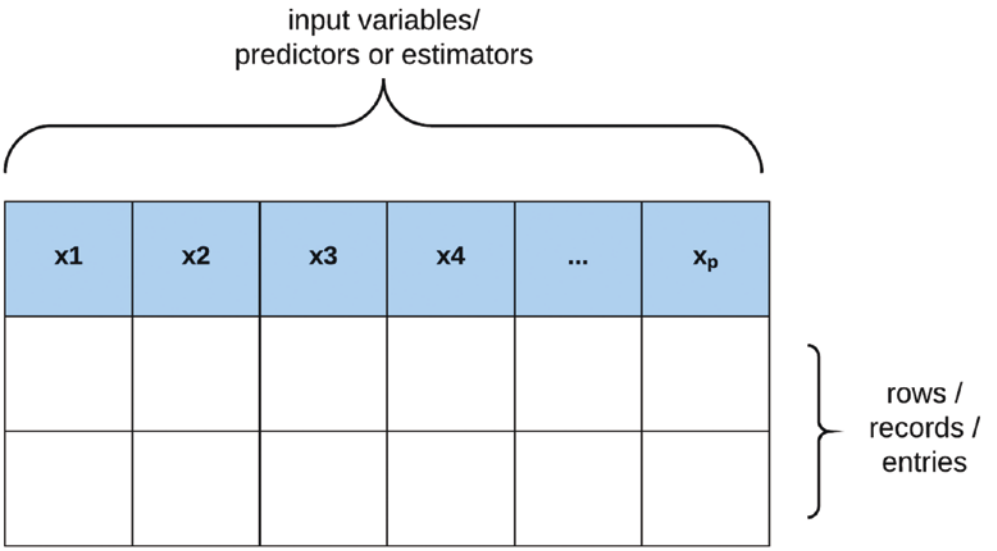

Figure 14-17. *Unsupervised dataset*

Rather, unsupervised learning algorithms attempt to determine the unknown structure of the dataset by grouping similar samples together.

Assume we have a dataset of patients with heart diseases; using unsupervised machine learning algorithms, we can find some hidden sub-groups of patients to help understand more about the disease patterns. This is known as *clustering*.

Also, we can use algorithms like *principal component analysis* (*PCA*) to compress a large number of features into principal components (that summarizes all the other features) for easy visualization. We will talk more about clustering and principal component analysis in later chapters.