# Unit 2 Glossary

## Array

A collection of numbers of a given type, such as float or int, in one or more dimensions.

## Axis

A particular dimension (or direction) in an array or a DataFrame.

## Binary indicators

A data transformation technique for transforming data to binary based on meeting a true/false condition.

## Bivariate plots

Plots that show a relationship between two data columns or two dimensions.

## Box-and-whisker plot (also known as box plot)

A type of data visualization to characterize the distribution of a set of numerical values, constructed by representing the following statistical quantities of the data set: median (50% level), first quartile (25%), third quartile (75%), minimum, and maximum.

## Categorical data type

A data type that is identified based on the label given to it. Two types are ordinal and nominal.

## Classification

The process of predicting categorical labels for previously unseen input data based upon prior training of a model (a "classifier") using labeled data examples.

## Classification model

A type of machine learning model for distinguishing among two or more discrete classes. For example, a natural language processing classification model could determine whether an input sentence was in French, Spanish, or Italian.

## Column

Each column in a data matrix contains the feature or attribute values. In supervised learning, one column contains an associated label value.

## Continuous data type

A data type consisting of standard floats or numbers; the distribution of continuous data types can produce outliers.

## DataFrame

A data table or spreadsheet with row and column headers, where each column contains data of a particular type but which can be of different types in different columns.

## Exploratory data analysis (EDA)

Looking at your data to answer three key questions: How is the data distributed? Which features are redundant? How do different features correlate with your label?

## Feature

An input variable used in making predictions. Features are the relevant characteristics or attributes of the data. For example, the features we might collect to identify fraudulent bank transactions might include dollar amount, type of transaction, country of origin, frequency, etc.

## Feature engineering

The process of determining which features might be useful in training a model then preparing and transforming the raw data into features that can be used for the model to train on.

## Functional transforms

A data transformation technique that transforms a numeric input X into a new numeric value based on f(X).

## Histogram

A type of data visualization to characterize the distribution of a set of numerical values, constructed by putting data into a set of discrete bins and plotting the number of counts in each bin.

## Interaction terms

A data transformation technique that takes the multiplication of two numeric types.

## Join

In database terminology (and carried over to related areas where one is working with multiple related data tables), the process of connecting different data tables together through a set of shared keys or labels. For example, in a company's employee database, there might be one table that stores employee personal information, one that stores work schedules, and one that stores

wage and salary information. They might all share an entry for each employee that encodes a unique employee ID, so one could join these different data tables by matching up employee IDs across tables.

## Label

What you want to predict. Your training data has labels so that you can train your model to predict the label of test examples. Each example in a labeled data set consists of one or more features and a label. For instance, in a housing data set, the features might include the number of bedrooms, the number of bathrooms, and the age of the house, while the label might be the house's price. In a spam detection data set, the features might include the subject line, the sender, and the email message itself, while the label would probably be either "spam" or "not spam."

## Labeled example

An example containing features and a label.

## Matrix

A 2-dimensional array of N vectors of size K.

## Nominative data type

A categorical type, often represented as strings. Nominative data types are the typical basis for classification.

## One hot encoding

The process of creating binary indicators from categorical value types. Each separate category would have its own binary indicator in one-hot-encoding.

## Ordinal data type

Consists of a mix of categorical and numeric. It is usually an integer-based number, but the range is usually so small that we can treat them as discrete categories. The absolute ordering matters in terms of being able to compare different values, but the relative difference doesn't.

## Row

In a data matrix, a row corresponds to an "example," also called a data point. In supervised learning, the row consists of features and one label.

## Sampling

The process of extracting subsets of examples from some available universe of data.

## Scatter plot

A type of data visualization for a pair of associated data sequences, where each pair is represented by a single point in the x-y plane — useful for visualizing the relationship between two sets of data values.

## Unit of analysis

A real-life representation of an example.

## Vector

A 1-dimensional array of K elements.

## Winsorization

The tranforsmation of statistics by setting extreme outliers equal to a specified percentile of the data to reduce the effect of having too many outliers.