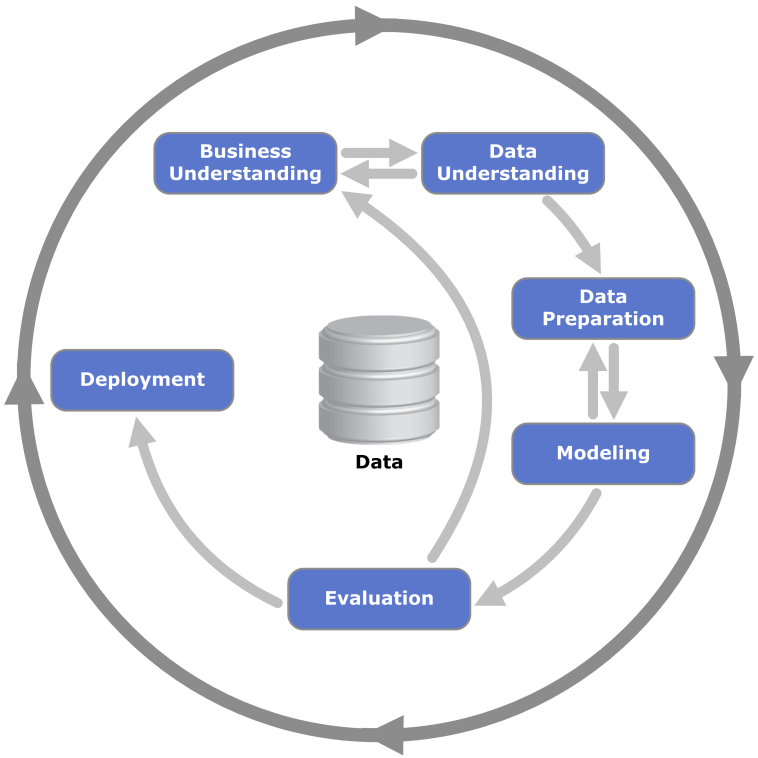
Machine learning

Managing Data



1. Problem Understanding lub Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Missing Values

There are several reasons why data could be missing.

- changes in data collection methods,
- human error,
- combining various datasets,
- human bias,
- and others.

Missing Values

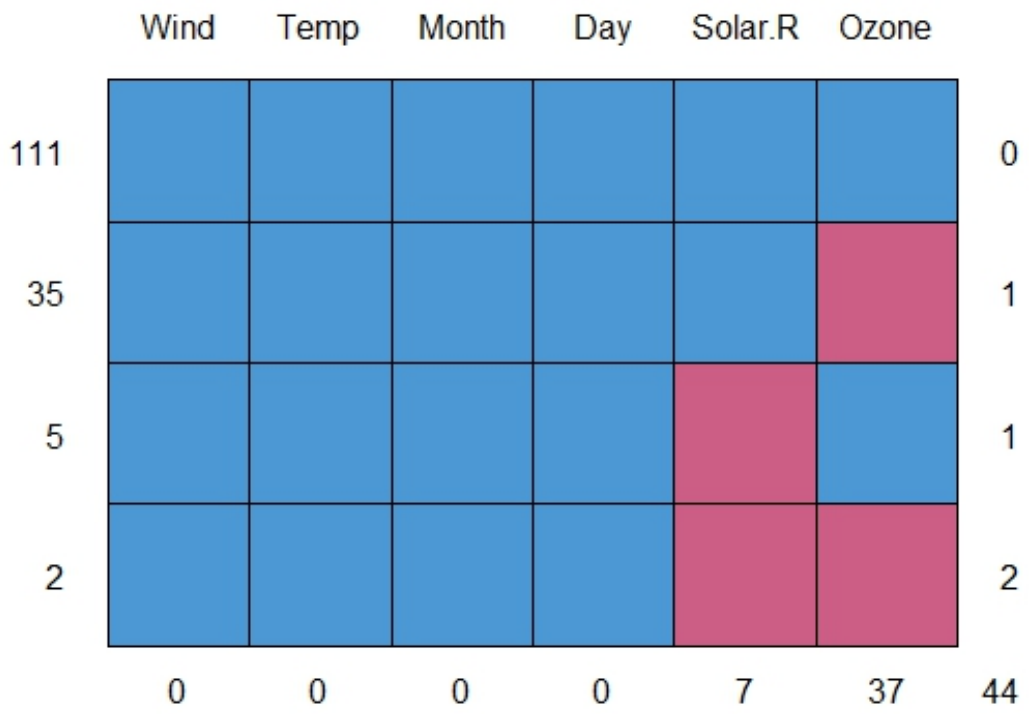
Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	78,000	C	M	5000
1002	J2S7K7	F	−40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

It is important to try to understand if there is a reason or pattern for the missing values.

For example, particular groups of people may not respond to certain questions in a survey.

Pattern for the missing values

- The **mice** package implements a method to deal with missing data.
- The mice package provides a function **md.pattern()** to get a better understanding of the pattern of missing data.



Dealing with missing data.

- **removal** - remove all instances with features that have a missing value.

- **this is a destructive approach and can result in the loss of valuable information and patterns that would have been useful in the machine learning process;**
- **this approach should be used only when the impact of removing the affected instances is relatively small or when all other approaches to dealing with missing data have been exhausted or are infeasible.**

Dealing with missing data.

- **imputation** - is the use of a systematic approach to fill in missing data using the most probable substitute values.
  - **random imputation** - involves the use of a randomly selected observed value as the substitute for a missing value. Disadvantage with this approach is that it ignores useful information or patterns in the data when selecting substitute values.
  - **distribution-based imputation** approach - the substitute value for a missing feature value is chosen based on the probability distribution of the observed values for the feature. This approach is often used for categorical values, where the **mode** for the feature is used as a substitute for the missing value.
  - **mean or median imputation** - involves the use of the mean or median of the observed values as a substitute for the missing value.
  - **predictive imputation** is the use of a predictive model (regression or classification) to predict the missing value. With this approach, the feature with the missing value is considered the dependent variable (class or response), while the other features are considered the independent variables.

Transforming the Data

As part of the data preparation process, it is often necessary to modify or transform the structure or characteristics of the data

- to meet the requirements of a particular machine learning approach,
- to enhance our ability to understand the data,
- to improve the efficiency of the machine learning process.

Z-score standardization

- **z-score, or zero mean normalization** - the approach results in normalized values that have a mean of 0 and a standard deviation of 1.

$$x^* = \frac{x - \bar{x}}{\sigma},$$

where x^* new value, x value from dataset.

Min-Max Normalization

- With **min-max normalization**, we transform the original data to a interval $[0, 1]$.

$$x^* = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)}.$$

```
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
```

Log Transformation

- The **logarithmic transformation** is applied to variables with distributions that are either not symmetric or have a wide range of values.

$$x^* = log(x_i).$$

Discretization

- **Discretization** involves treating continuous features as if they are categorical.
- This is often done as a pre-step before using a dataset to train a model. This is because some algorithms require the independent data to be binary or to have a limited number of distinct values.

Dummy coding involves the use of numeric values to represent categorical features. Dummy coding is often used for algorithms that require that the independent features be numeric.

Dummy coding

Drive	Code
Front-Wheel Drive	1
Rear-Wheel Drive	2
All-Wheel Drive	3

Drive	Front-Wheel Drive	Rear-Wheel Drive	All-Wheel Drive
Front-Wheel Drive	1	0	0
Rear-Wheel Drive	0	1	0
All-Wheel Drive	0	0	1

Drive	Front-Wheel Drive	Rear-Wheel Drive
Front-Wheel Drive	1	0
Rear-Wheel Drive	0	1
All-Wheel Drive	0	0

Thank you for your attention!!!