Feature Engineering

Raw data
Features

INTRODUCTION TO FEATURE ENGINEERING

- Feature Engineering encapsulates various data engineering techniques such as selecting relevant features, handling missing data, encoding the data, and normalizing it.

- It can be thought of as the art of selecting the important features and transforming them into refined and meaningful features that suit the needs of the model.

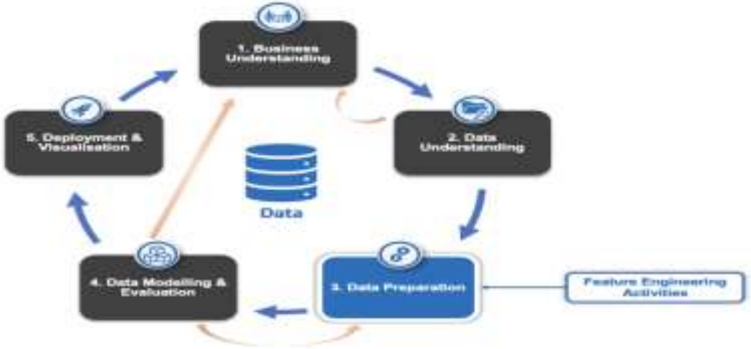
- Feature Engineering is the process of extracting and organizing the important features from raw data in such a way that it fits the purpose of the machine learning model.

- It is one of the most crucial tasks and plays a major role in determining the outcome of a model.
- Feature Engineering consists of Preprocessing and Feature Selection

IMPORTANCE OF FEATURE ENGINEERING

- Feature engineering is focused on using the variables you already have to create additional features that are (hopefully) better at representing the underlying structure of your data.Feature engineering is a creative process that relies heavily on domain knowledge and the thorough exploration of your data.

- After the data is cleaned and processed it is then ready to be fed into the machine learning models to train and generate outputs.A feature is a variable that is important for predicting your specific target and addressing your specific question.


1. Business Understanding
2. Data Understanding
3. Data Preparation
Feature Engineering Activities
4. Data Modelling & Evaluation
5. Deployment & Visualisation
Data

NULL VALUES

- Null values, also known as missing values, represent the absence of data in a particular field or variable within a dataset.Null values can occur for various reasons, such as data entry errors, data collection limitations, or intentional absence of information.

- A null value is a placeholder that indicates the absence of a value.Null values exist for all data types. The null value of a given data type is different from all non-null values of the same data type.

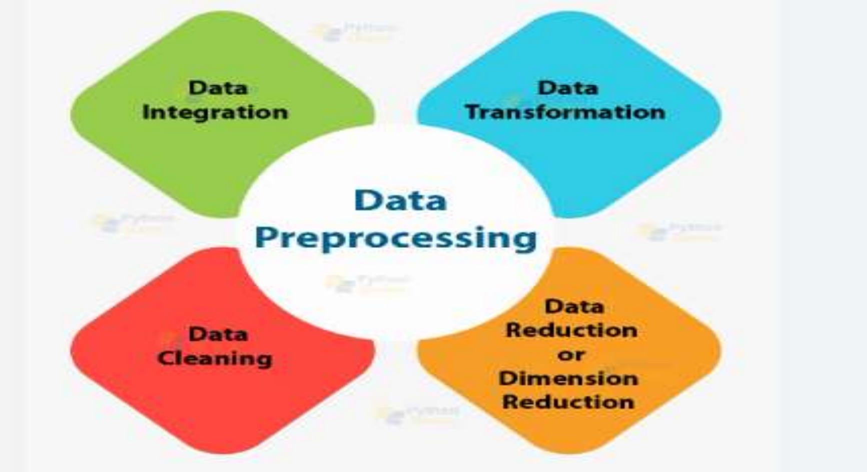
- isnull() defines missing values in a series or Dataframe.

- not null() check for missing values in a pandas Series or DataFrame. It returns a boolean Series or DataFrame, where True indicates non-missing values and False indicates missing values.

PREPROCESSING

- Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.It is the first and crucial step while creating a machine learning model.

- Preprocessing often involves cleaning, transforming, and organizing the data to make it more useful and meaningful for subsequent analysis.



IMPORTANCE OF DATA PREPROCESSING

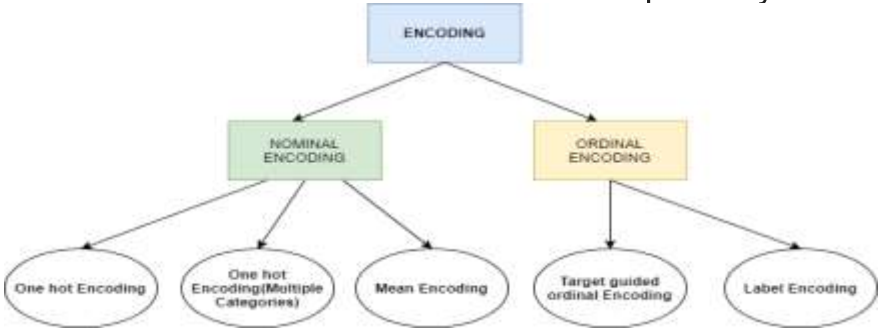
- The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis.

- Feature engineering practices that involve data wrangling, data transformation, data reduction, feature selection and feature scaling help restructure raw data into a form suited for particular types of algorithms.

ENCODING TECHNIQUES

- It refers to the process of converting categorical or textual data into numerical format, so that it can be used as input for algorithms to process.The reason for encoding is that most machine learning algorithms work with numbers and not with text or categorical variables.

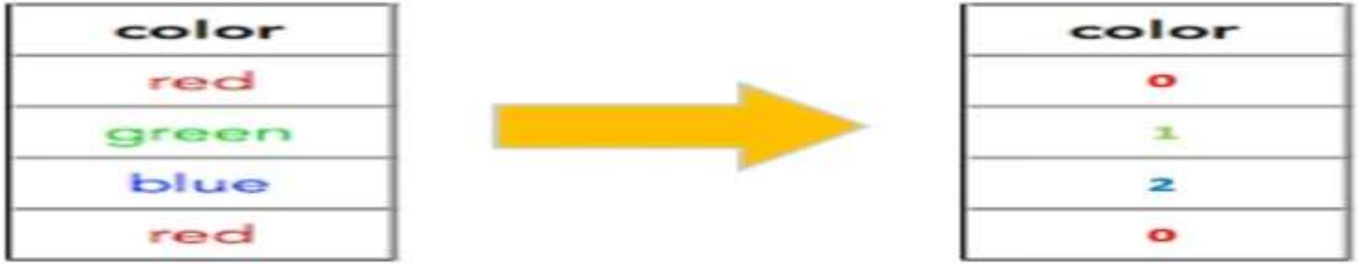
- Encoding categorical variables is a vital step in preparing data for machine learning tasks.When dealing with categorical data, characterized by non-numeric values such as text or categories, it becomes necessary to transform them into a numerical format for compatibility with machine learning algorithms.



LABEL ENCODING

- Label encoding is a technique used to convert categorical data into numerical format by assigning a unique integer value to each category or label in a feature.

- Label encoding is suitable for categorical features with only two distinct categories.It is a simple and effective way to prepare categorical data for machine learning algorithms that require numerical input.



color
red
green
blue
red

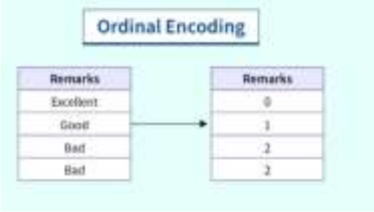
color
0
1
2
0

ORDINAL ENCODING

- Ordinal encoding is a technique used to convert categorical data into numerical format while preserving the ordinal relationship among the categories.

- Ordinal encoding is similar to label encoding but allows you to explicitly define the mapping between categories and integer labels.

- Ordinal encoding is similar to label encoding but allows you to explicitly define the mapping between categories and integer labels.This is especially useful when there is a clear and predefined ordinal relationship.



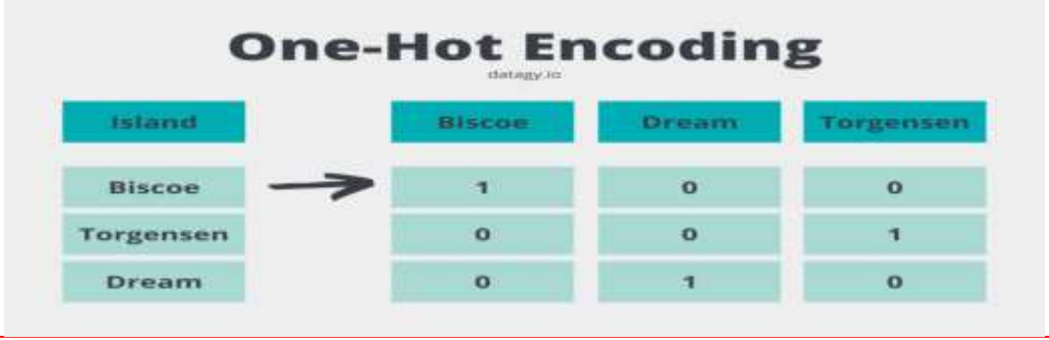
Remarks
Excellent
Good
Bad
Bad

Remarks
0
1
2
2

ONE-HOT ENCODING

- One-hot encoding is a method of converting categorical data into a numerical format by representing each category as a binary vector.It is suitable for nominal categorical variables, where the categories have no inherent order or relationship.

- In a one-hot encoded vector, each category is assigned a unique index, and the corresponding element in the vector is set to 1, while all other elements are set to 0.



One-Hot Encoding

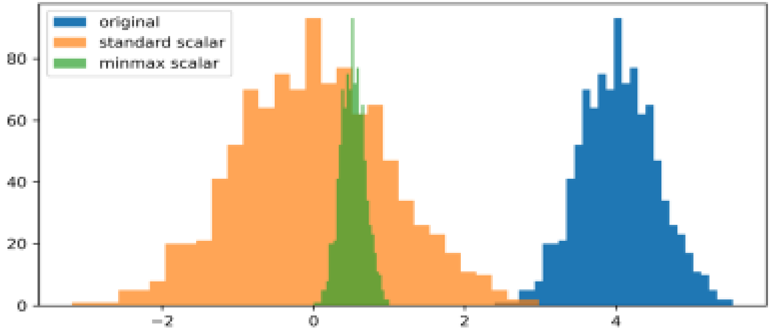
datagy.io

Island		Biscoe	Dream	Torgensen
Biscoe	→	1	0	0
Torgensen		0	0	1
Dream		0	1	0

SCALING

- Scaling is a data preprocessing technique used to standardize or normalize the numerical features of a dataset to a similar scale. Feature scaling is the process of normalizing the range of features in a dataset.

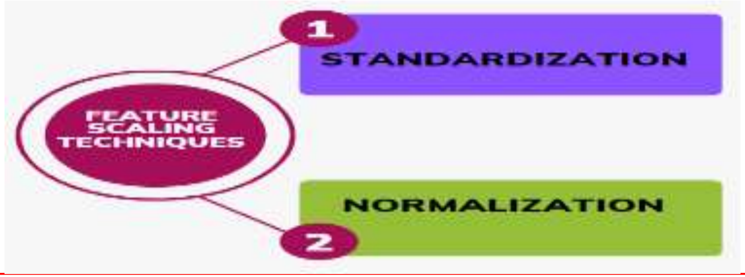
- Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.



IMPORATANCE OF SCALING

- Scaling the data can help to balance the impact of all variables on the distance calculation and can help to improve the performance of the algorithm.Now it’s time to see why scaling is important before model building or how it can improve the model’s accuracy.

- Scaling the target value is a good idea in regression modelling; scaling of the data makes it easy for a model to learn and understand the problem.



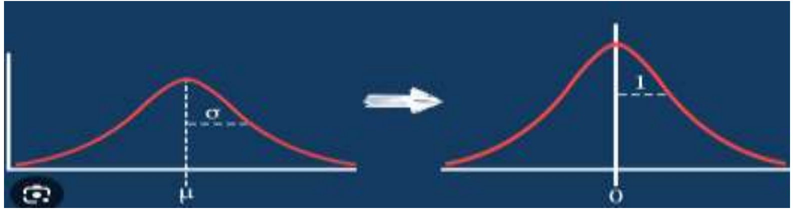
STANDARDIZATION

- The basic concept behind the standardization function is to make data points centred about the mean of all the data points presented in a feature with a unit standard deviation.This means the mean of the data point will be zero and the standard deviation will be 1.

- This technique also tries to scale the data point between zero to one but in it, we don’t use max or minimum.In statistics, the mean is the average value of all the numbers presented in a set of numbers and the standard deviation is a measurement of the dispersion of the data points from the mean value of the data points.

- So in standardization, the data points are rescaled by ensuring that after scaling they will be in a curve shape. Mathematically we can represent it as follow :

$$X(\text{std}) = X - \text{mean}(X) / \text{std deviation}(X).$$



NORMALIZATION

- Normalization can have various meanings, in the simplest case normalization means adjusting all the values measured in the different scales, in a common scale.It is a very common approach to scaling the data.
- In this method of scaling the data, the minimum value of any feature gets converted into 0 and the maximum value of the feature gets converted into 1.
- We can represent the normalization as follows :

$$X(norm) = X- \min(x) / \max(x) - \min(x)$$

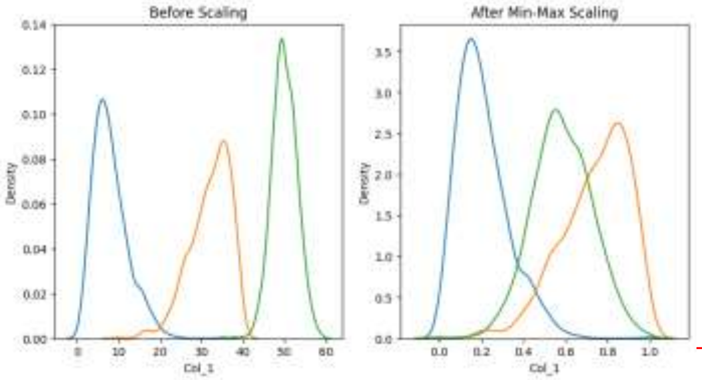
- Where x is any value from the feature x and min(X) is the minimum value from the feature and max(x) is the maximum value of the feature.

0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
z-score

MIN_MAX SCALING

- In many scaling techniques, such as Min-Max Scaling or Standardization, the minimum value of a feature is a key factor in determining the transformation applied to the data.In Min-Max Scaling, the minimum value of the feature is subtracted from each data point before scaling.

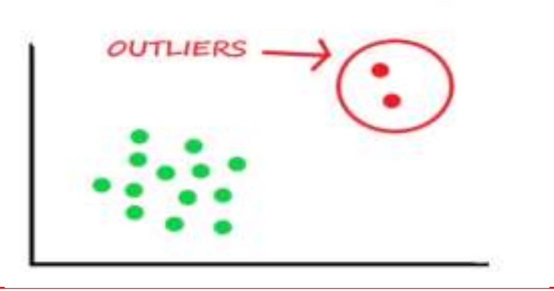
- This ensures that the minimum value of the feature in the original dataset is mapped to 0 after scaling.



OUTLIERS

- An outlier is an observation point that is distant from other observations.

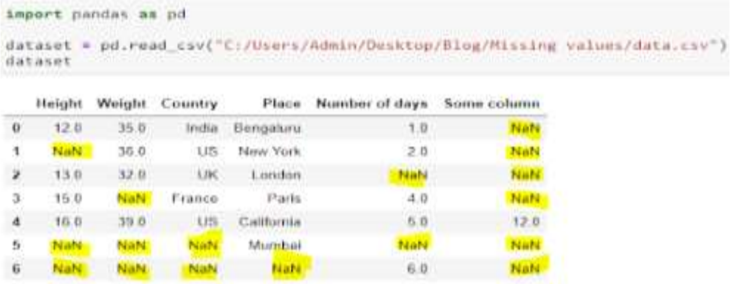
- An outlier may be due to variability in the measurement or it may indicate experimental error;the latter are sometimes excluded from the data set.



MISSING VALUES

- Missing values are data points that are absent for a specific variable in a dataset.They can be represented in various ways, such as blank cells, null values, or special symbols like “NA” or “unknown.”

- These missing data points pose a significant challenge in data analysis and can lead to inaccurate or biased results.Handling missing values is a critical aspect of data preprocessing in data analysis and machine learning tasks, as they can introduce biases, affect statistical analyses, and lead to incorrect conclusions if not addressed properly.



```
import pandas as pd

dataset = pd.read_csv("C:/Users/Admin/Desktop/Blog/Missing values/data.csv")
dataset
```

	Height	Weight	Country	Place	Number of days	Some column
0	12.0	35.0	India	Bengaluru	1.0	NaN
1	NaN	36.0	US	New York	2.0	NaN
2	13.0	32.0	UK	London	NaN	NaN
3	15.0	NaN	France	Paris	4.0	NaN
4	16.0	39.0	US	California	5.0	12.0
5	NaN	NaN	NaN	Mumbai	NaN	NaN
6	NaN	NaN	NaN	NaN	6.0	NaN

TRANSFORMERS

- Transformers refer to objects or functions that modify or transform data in some way. These transformations are applied to prepare raw data for downstream tasks such as modeling, analysis, or visualization.

- Transformers play a crucial role in data preprocessing pipelines, where they clean, preprocess, and engineer features to make the data suitable for machine learning algorithms.

DATA TRANSFORMATIONS

> LOG TRANSFORMER
> RECIPROCAL TRANSFORMER
> SQUARE TRANSFORMER
> SQUARE ROOT TRANSFORMER
> BOX-COX TRNASFORMER

LOGARITHMIC TRANSFORMATION(log)

- A logarithmic transformation is a mathematical operation that involves taking the logarithm of each data point in a dataset.It is commonly used in data analysis and statistical modeling to address issues such as skewed distributions, heteroscedasticity, or to linearize relationships that are exponential in nature.

- Applies the natural logarithm (ln) to a feature (column) in your dataset.
- Useful for right-skewed data (data concentrated towards higher values).
- Makes the distribution more symmetrical and reduces the impact of outliers.

Example (assuming `x` is your feature):

```
transformed_x = np.log(x)
```

SQUARE ROOT TRANSFORMATION(sqrt)

- Square root transformation is a data preprocessing technique that involves taking the square root of each data point in a dataset.

- The square root transformation is often applied to continuous variables with skewed distributions, such as count data or data with a Poisson distribution. By taking the square root of each data point, the transformation reduces the impact of extreme values and makes the data distribution closer to normality.

- Suitable for data with positive values that are heavily skewed to the right.Similar to the log transformation, it compresses the larger values and spreads out the smaller ones.

```
transformed_x = np.sqrt(x)
```

BOX-COX TRANSFORMATION

- The Box-Cox transformation is a data transformation technique used to stabilize variance and make the data more closely approximate a normal distribution.

- It is particularly useful when dealing with data that violates the assumptions of normality and homoscedasticity (constant variance) required by many statistical models.

- The Box-Cox transformation can improve the performance of statistical models that assume normally distributed residuals, such as linear regression and ANOVA.

THANK
YOU !