Basic Machine Learning: Data Preprocessing

Goal

Goal

To understand how important data preprocessing is in term of building machine learning model

- Understand & prepare the datasets for machine learning modeling process
- Important task technique that transforms raw data into a more understandable, useful and efficient format

Outline

Outline

- Data Handling

- Data Transformation
  - ❑ Categorical data transformation
  - ❑ Numerical data transformation

- Feature Analysis
  - ❑ Dimensional reduction & TDA
  - ❑ Feature analysis
  - ❑ Feature engineering

Content

Data Handling

Our real world data is generally:

Missing or incomplete: Certain attributes or values or both are missing or only aggregate data is available

Noisy: Data consists of errors, outliers, inconsistency

Categorical or Continuous: Data represents categorical or continuous values that are not standardized

Data Handling

How to deal with such data?

- Replace with default value
- Replace with mean/median/mode values
- Drop data

Which one to use?

Data Transformation

To transform data into form that can be learned easier by computer

To simplify data by decreasing its value scale

All based on whether the data type is categorical or numerical

Data Transformation

To transform data into form that can be learned easier by computer

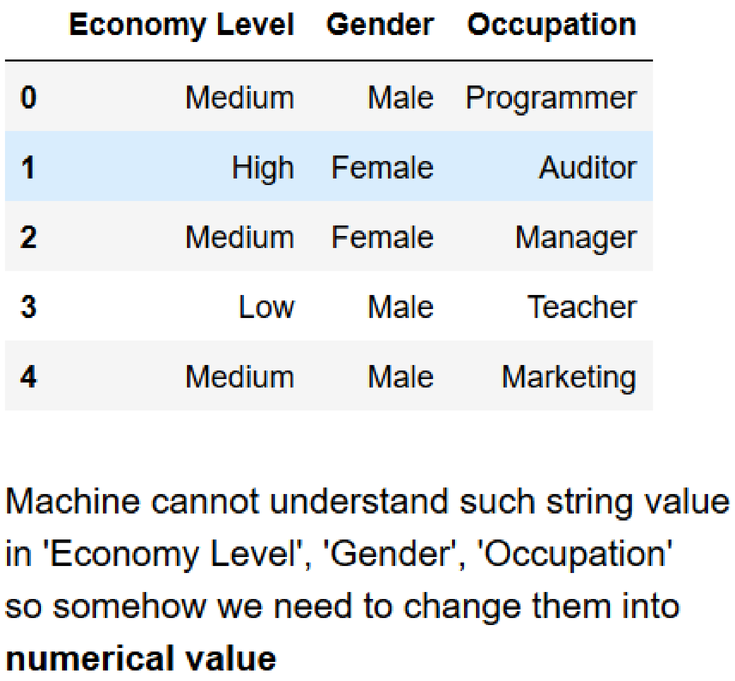
To simplify data by decreasing its value scale

To standardize data type (categorical or numerical)

Data Transformation (Categorical)

The idea is to make all features that contain categorical data for having numerical information. Generally there are 2 method that can be used:

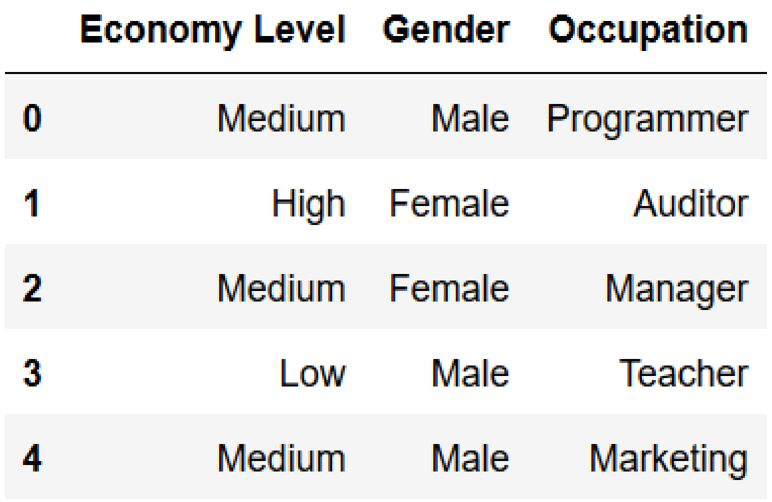
1. Label Encoder
2. One Hot Encoder



	Economy Level	Gender	Occupation
0	Medium	Male	Programmer
1	High	Female	Auditor
2	Medium	Female	Manager
3	Low	Male	Teacher
4	Medium	Male	Marketing

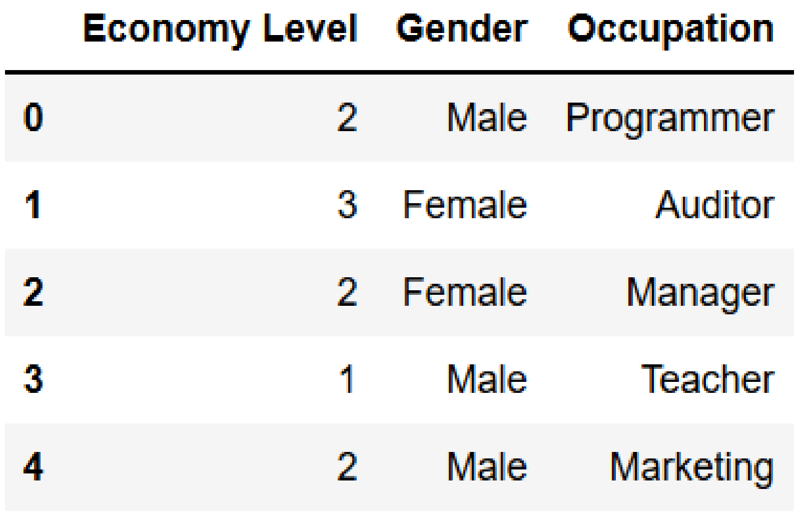
Machine cannot understand such string value in 'Economy Level', 'Gender', 'Occupation' so somehow we need to change them into **numerical value**

Data Transformation (Categorical)



	Economy Level	Gender	Occupation
0	Medium	Male	Programmer
1	High	Female	Auditor
2	Medium	Female	Manager
3	Low	Male	Teacher
4	Medium	Male	Marketing

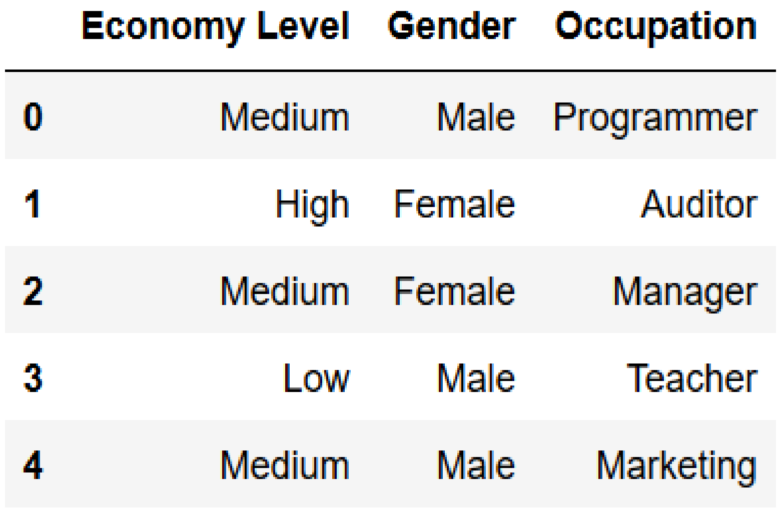
⇨



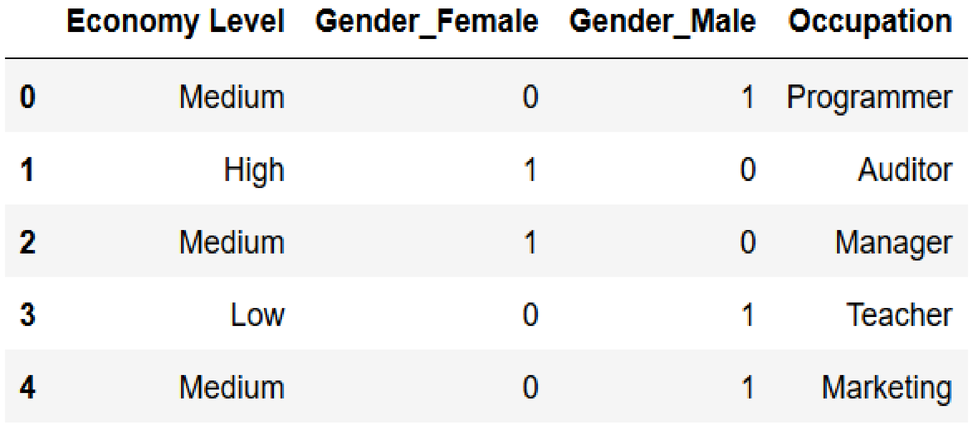
	Economy Level	Gender	Occupation
0	2	Male	Programmer
1	3	Female	Auditor
2	2	Female	Manager
3	1	Male	Teacher
4	2	Male	Marketing

Label Encoder

Data Transformation (Categorical)



	Economy Level	Gender	Occupation
0	Medium	Male	Programmer
1	High	Female	Auditor
2	Medium	Female	Manager
3	Low	Male	Teacher
4	Medium	Male	Marketing



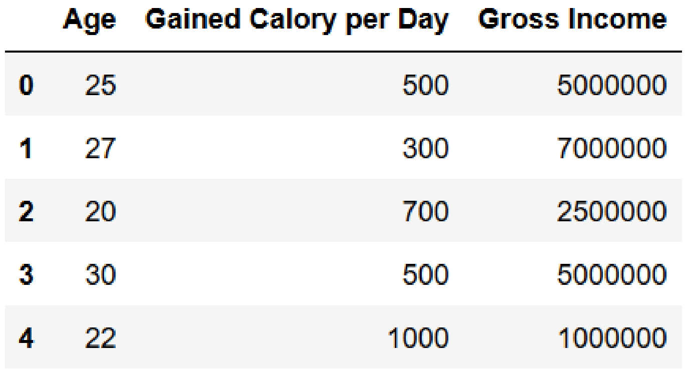
	Economy Level	Gender_Female	Gender_Male	Occupation
0	Medium	0	1	Programmer
1	High	1	0	Auditor
2	Medium	1	0	Manager
3	Low	0	1	Teacher
4	Medium	0	1	Marketing

One Hot Encoder

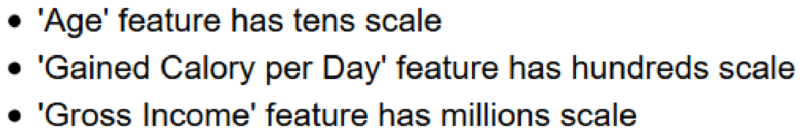
Data Transformation (Numerical)

The idea is to make all feature that contain numerical data for having same scale. Generally there are 2 method that can be used:

1. Standard Scaler
2. Min-Max Scaler

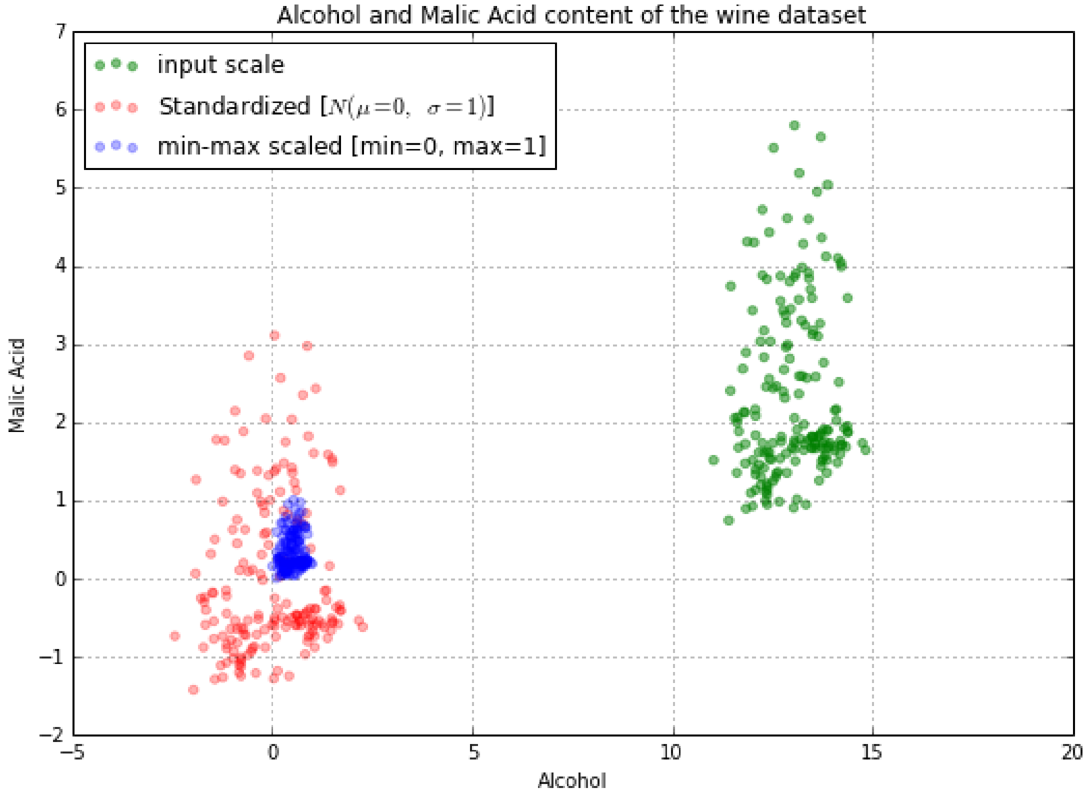


	Age	Gained Calory per Day	Gross Income
0	25	500	5000000
1	27	300	7000000
2	20	700	2500000
3	30	500	5000000
4	22	1000	1000000



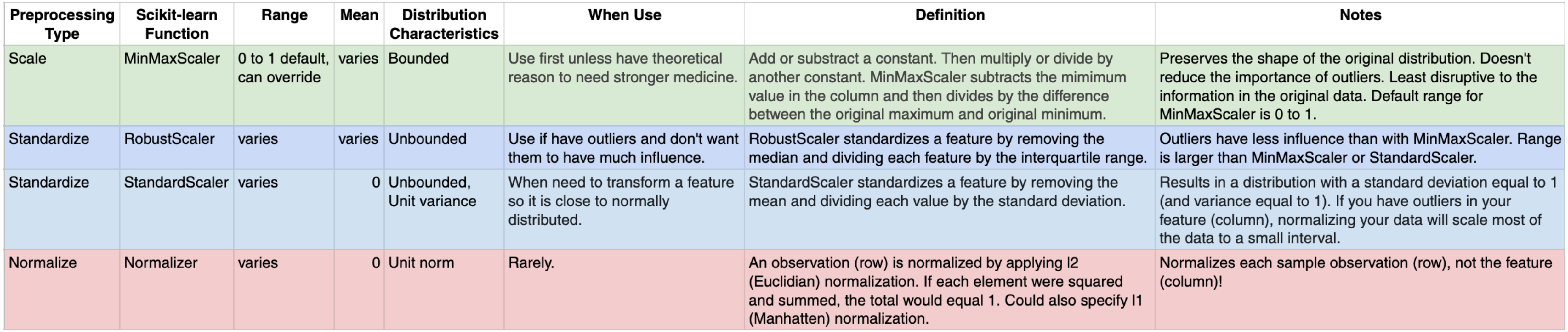
- 'Age' feature has tens scale
- 'Gained Calory per Day' feature has hundreds scale
- 'Gross Income' feature has millions scale

Data Transformation (Numerical)



- Standard Scaler transform data distribution into **normal distribution**
- Min Max Scaler transform data range **into 0 ~ 1**
- Rule of thumb:
  - Use Min Max Scaler as the default if you are transforming a feature
  - Use Standard Scaler if you need a relatively normal distribution

Data Transformation (Numerical)



Preprocessing Type	Scikit-learn Function	Range	Mean	Distribution Characteristics	When Use	Definition	Notes
Scale	MinMaxScaler	0 to 1 default, can override	varies	Bounded	Use first unless have theoretical reason to need stronger medicine.	Add or substract a constant. Then multiply or divide by another constant. MinMaxScaler subtracts the mimimum value in the column and then divides by the difference between the original maximum and original minimum.	Preserves the shape of the original distribution. Doesn't reduce the importance of outliers. Least disruptive to the information in the original data. Default range for MinMaxScaler is 0 to 1.
Standardize	RobustScaler	varies	varies	Unbounded	Use if have outliers and don't want them to have much influence.	RobustScaler standardizes a feature by removing the median and dividing each feature by the interquartile range.	Outliers have less influence than with MinMaxScaler. Range is larger than MinMaxScaler or StandardScaler.
Standardize	StandardScaler	varies	0	Unbounded, Unit variance	When need to transform a feature so it is close to normally distributed.	StandardScaler standardizes a feature by removing the mean and dividing each value by the standard deviation.	Results in a distribution with a standard deviation equal to 1 (and variance equal to 1). If you have outliers in your feature (column), normalizing your data will scale most of the data to a small interval.
Normalize	Normalizer	varies	0	Unit norm	Rarely.	An observation (row) is normalized by applying l2 (Euclidian) normalization. If each element were squared and summed, the total would equal 1. Could also specify l1 (Manhatten) normalization.	Normalizes each sample observation (row), not the feature (column)!

Feature Analysis

Feature Selection	Feature Engineering	Feature Extraction
▪ Based on domain knowledge ▪ Exclude unnecessary features ▪ Heatmap and many other visualization techniques support	▪ Based on domain knowledge ▪ Generating new feature from other related datasets	▪ Keep all information from all feature before extracting the information ▪ Popular techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Auto Encoder (Neural Network Approach)

Thanks!