Data Science With Python: Data Pre-processing #4747

Name: Deepthi M

Serial Number: 172

Batch Number: 05

Data Pre-processing

- Data Pre-processing is one of the most important steps in data science and these steps require the most time in a project. It is an important step in any project.

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

- There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Data cleaning involves the following steps:

i. Typecasting

ii. Handling Duplicates

iii. Outlier Analysis / Treatment

iv. Zero & Near Zero Variance Features

v. Missing Values

vi. Discretization / Binning / Grouping

vii. Dummy Variable Creation

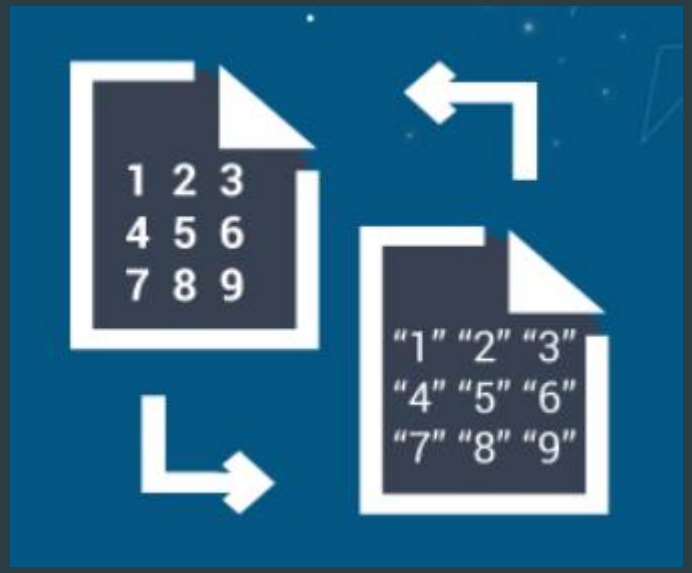
viii. Normalization / Standardization

ix. Transformation

x. String Manipulations (Unstructured Textual Data)

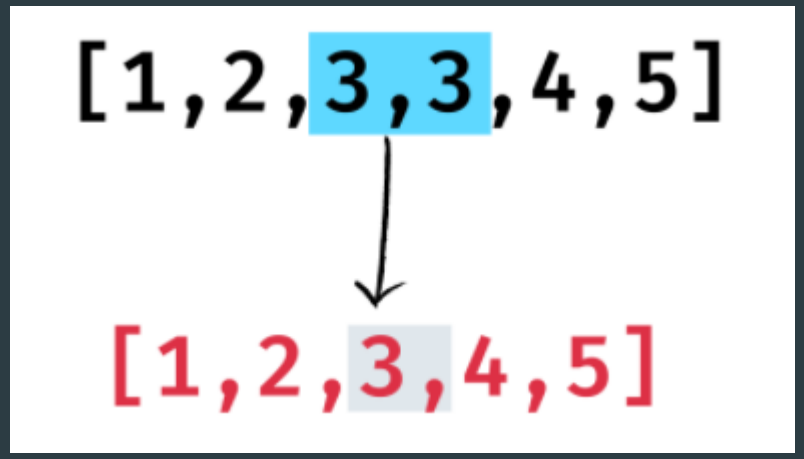
Typecasting:

- Conversation of one datatype variable to another.
- Typecasting is important as it helps in converting the data types of one variable to another.
- Example, in deep learning the weights are assigned are float in nature. Hence, it is necessary to change the data type of the input variables to float.



Handling Duplicates:

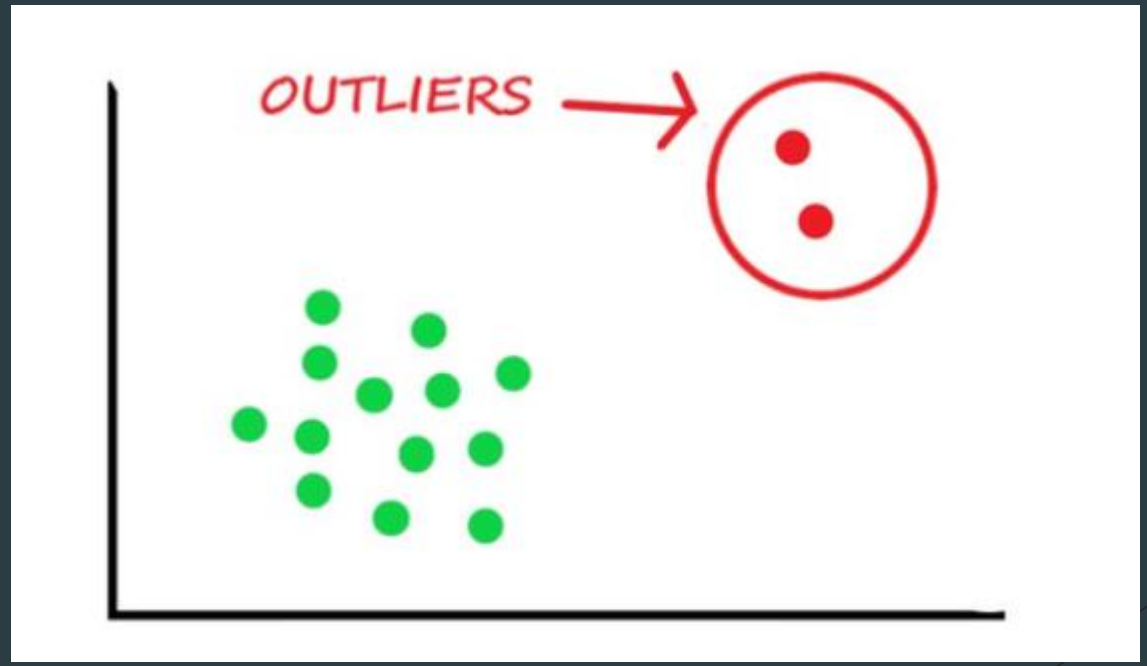
- Duplicated values or repeated values need to be removed from the dataset.
- May impact the model as the values perform or contribute the same for the model, which might impact the model performance.



Outlier Analysis / Treatment:

- Extreme values in each column.
- Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner.
- Impact on accuracy is more when we don’t treat outliers, hence conversion or removing is necessary as it helps in better performance of the model.

- Steps to treat the outliers:
- Winsorization: Converting the extreme values to the nearest values.
- Trimming: Cutting off the extreme values.



Zero & Near Zero Variance Features:

- Values in the columns having variance zero or near zero.
- Columns having zero or near-zero variance are usually avoided because applying mathematical operations on these columns may not impact much on models as they are having zero or near-zero variance.

Missing Values:

- Rows in the columns when are nan values or empty known as missing values.

- Imputation methods are used to deal with missing values.
- Typically, they ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values.

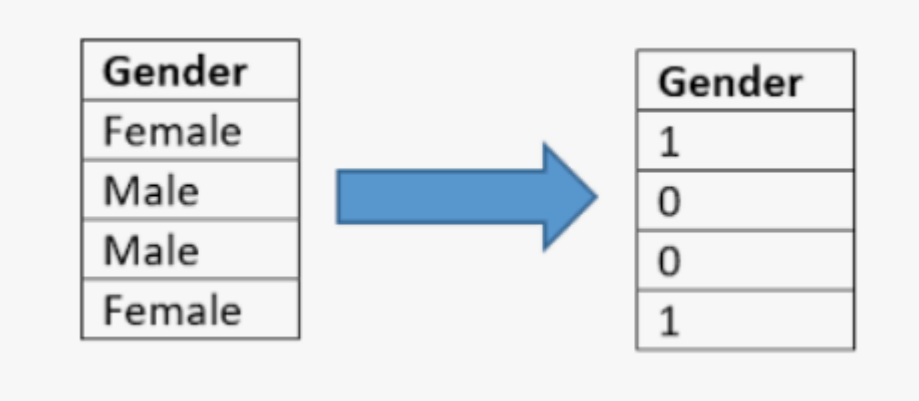
Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+		80%
5	Lite	76	70%
6	Fast+	155	10%
7			95%
8	Lite	76	77%
9	Fast+	180	

Discretization / Binning / Grouping:

- Conversation of continuous values to discrete values.

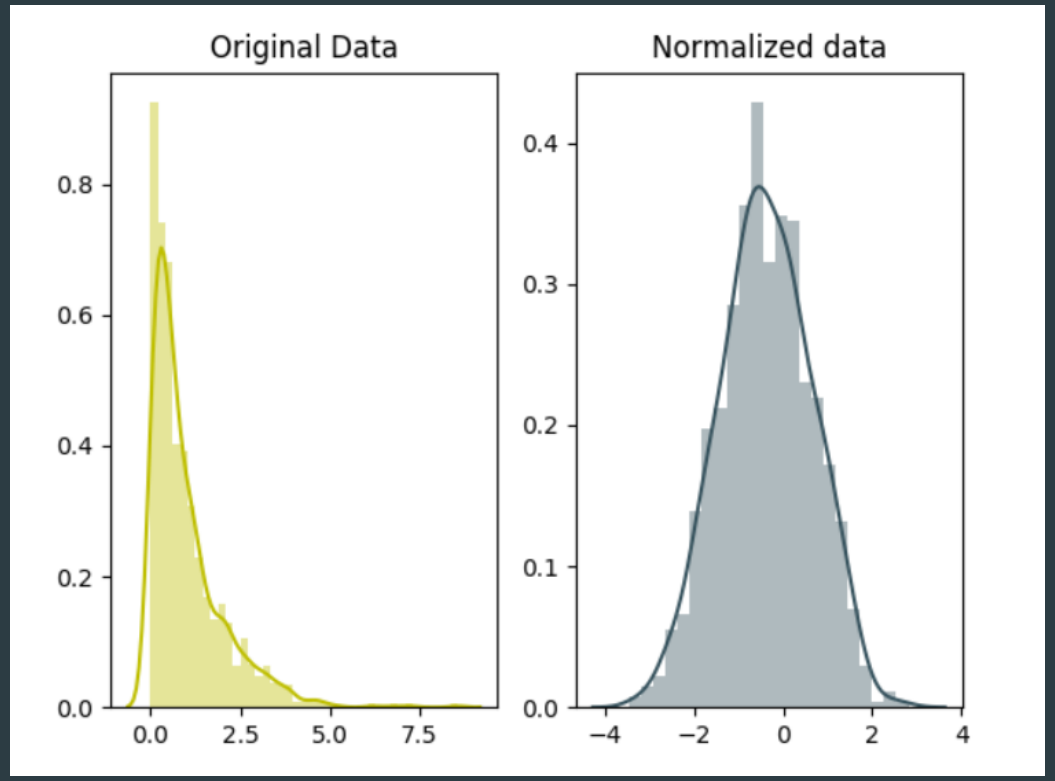
Dummy Variable Creation:

- Conversation of categorical data to numeric data.
- Many of the models in data mining treats it computes the dataset when all the data type is numeric.
- Methods that are used to convert:
- Label encoder: Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.
- One- hot encoding: A one-hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values.



Normalization / Standardization:

- Converting all the values in the data to the same distribution.
- It helps in converting data to unit less or scale free.
- It also helps in increasing the model performance and accuracy.



Advantages:

- It helps to increase the model training.
- Increases efficiency.
- Reduces the complexity of the dataset.
- Reduces the training time.
- The model performs better than before.

Disadvantages:

- Losses the original values in the dataset.
- Limited to numeric data.