Based on the Exploratory Data Analysis (EDA)of the data, can we do preprocessing /Feature Engineering of the data.

Next step in Data science Life Cycle is **Preprocessing / Feature Engineering (FE)**

In the following areas we can do preprocessing of the data.

* Handling missing values
* Outlier handling
* Scaling the data
* Transformation of the data (Log transformation, BoxCox Transformation, Square, Cube transformation etc)
* Encoding
* Duplicate values handling
* Imbalanced data handling
* Feature Selection
* Dimensionality Reduction etc
* Split/Merge/drop/add data

Let us go little more deep into explanation.

**Ways of preprocessing /FE**

**Missing value handling:**

* Randomly filling the missing values
* Forward filling backward filling
* Statistical approach (Mean/Median/Mode)
* End of the distribution
* Drop that row
* KNN –imputer
* Create our own ML model to predict the missing values

**Outlier Handling:**

|  |  |
| --- | --- |
| **Detect the Outlier** | **Handling the outlier** |
| Z-score | By dropping |
| IQR | By filling with median |
| Box Plot | Replace / trimming with any value |
| Scatter Plot |  |
| Violin Plot |  |

**Transformation of data:**

* Box-Cox transformation
* Power Transformation
* Log Transformation
* Square Transformation
* Cube Transformation
* Yeo Johnson transformation

**Scaling of the data:**

* Standardization
* Min-max scaler
* Unit scaling

**Encoding**:

* One hot encoding
* Label encoding
* Binary encoding
* Target guided encoding
* Hash encoding

**Imbalanced data handling:**

* By collecting more data
* Under sampling
* Oversampling
* Cluster based over sampling

After completion of Feature engineering, now our data become very clean, without missing values or outliers etc.

Now it’s time for us to **build our own Model**.

Will explain you the model building with practical implementation (Complete code from scratch).

There we will be covering **Model Evaluation / Validation**.

Based on respective **performance metrics**, we can recommend the best model to predict the output.