**Feature Scaling (FE)**

**What is One Hot Encoding:** whenever we have nominal categorical data then we split the categories of that feature into columns i.e against each category we create a column and assign them 1 and 0.

**Note:** multiclonality is process in which we remove a column from those which are created with the help of one hot encoding.

**Note:** when ever we got lot numbers of dimensions due to one hot encoding then what we do is. We only keep most frequent categories and convert less frequent categories into a new single category.\

**Pipelines**

Pipelines chains together multiple steps so that the output of each step is used as input to the next step. Pipelines makes it easy to apply the same preprocessing to train and test.

Always use pipeline because it is very useful in the prospective of production of the software. Code in (Feature scaling folder)

**Note:** Project life cycle = (**[Data Preprocessing]**EDA -> Feature Engineering-> Feature Scaling)->(**[Model Training]**Model creation) -> model deployment

* **Data preprocessing** has object of Transformers: which are used to for feature engineering. i.e scaling the feature e.g StandardScaler, MinMaxScaler which transform the feature mean = 0 and standard deviation = 1.
* **Model Training** has object of Model: which are used to train the model. E.g Linear Regression, Logistic Regression, Decision Tree.

**Note: fit(), transform() , fit\_transform() and predict() methods in scikit learn differences.**

|  |  |
| --- | --- |
| **Transformer** | **Model** |
| For training data we do fit and transform | for training data we do fit |
| For Test data we do transform only | For test data we do predict |

**Fit ():** when we apply fit method on training data it actually only computes the standardization whose formula is = Xi – mean / standard deviation

**Transform ():** when we apply transform it actually apply the standardization on whole of the feature values. And there mean = 0, and SD = 1.

**Fit\_tranform():** does both the things at same time.

**Predict ():** is used with model to predict on test or new data.

**Feature Transformation (FE)**

1. Log transform
2. Reciprocal transform
3. Square root transform

**Note**: What happens after applying the transformation? Why your model improves?

Ans is that the Probability Density (PD) function is converted into normal distribution.