**Data Pre-processing**

1. Handling missing data
2. If amount of missing data is way less than complete data, deleting those rows is a go-to idea.
3. If not, then replace the missing data with average of all data in that column.(Recommended)
4. Replace with Median salary
5. Replace with most frequent

Make use of SimpleImputer function in sklearn library

1. Encoding Categorical Data

One hot encoding :- creating binary vector of distinguished category (not recommended for just true/false or yes/no value… use 0,1 instead)

1. Encoding Dependent Variable

LabelEncoder : - for 0 and 1 type of encoding

* Feature Scaling should be done after splitting dataset into training and test set, because test set is going to be a brand new dataset which we are not going to work with, if feature scaling is applied before splitting, it’ll provide information leakage before training is done on test set by giving mean, median ,standard deviation etc info before training is done.

1. Feature Scaling

* To avoid some features to be dominated by some other features in such a way that dominated features are not considered by ML model.
* No need to apply for all ML model
* Only to be applied in training set, not test set
* Not to be applied on dummy/ encoded variables , we’ll lose interpretation of the data

2 techniques :-

1. Standardisation – works always

Xstand = (X-mean(X))/standard deviation(X)

1. Normalisation – recommended when normal distribution is present

Xnorm = (X – min(X))/(max(X) – min(X))

**Simple Linear Regression**

Y= b0 + b1x

1. Dependent and Independent Variable

Y in above equation is dependent variable and X is Independent Variable. B is a coefficient which is a relation constant of X and Y.

1. How to find Best Fitting Line

Find minimum sum of squares of difference to draw a straight line.

Steps for SLR :

1. Importing the libraries
2. Importing the dataset
3. Splitting the dataset into the Training set and Test set
4. Training the Simple Linear Regression model on the Training set
5. Predicting the Test set results
6. Visualising the Training set results
7. Visualising the Test set results

**Multiple Linear Regression**

Y= b0 + b1x1 + b2x2 + ….. + bnxn

Dummy Variables :- For categorical data

No Need to apply Feature scaling

Steps :-

* Import the libraries
* Import the dataset
* Encoding categorical data
* Splitting the dataset into the Training set and Test set
* Training the multiple linear regression model on training set
* Predicting the test results

**Building a Model**

5 methods of model building :

1. All-In
2. Backward Elimination
3. Forward Elimination Stepwise Regression
4. Bi-directional Elimination
5. Score Comparison

* All-In :- Using all Variables to build a model
* All possible Models :-
* Step 1 – Select a criteria for of goodness of fit ( e.g Akaike criteria)
* Step 2 – Construct all possible regression models. 2N -1 total combinations.
* Step 3 – Select the one with best criteria
* Finish – Your model is ready
* Forward – Elimination :-
* Step 1 – Select a Significance Level to enter the model (e.g SL =0.05)
* Step 2 – Fit all simple Regression model y ~ xn. Select one with lowest p-value.
* Step 3 – Keep this variable and fit all other models with one extra predictor added to the one(s) we already have.
* Step 4 – Consider the predictor with lowest p- value. If p< Significance Level. Go to Step 3 otherwise go to finish.
* Finish – Choose the previous model
* Backward-Elimination :-
* Step 1 - Select a significance level to stay in the model
* Step 2 – Fit the full model model with all possible predictors
* Step 3 – Consider the predictor with highest p-value. If p > Significance Level then go to Step 4 else go to finish.
* Step 4 – Remove the predictor
* Step 5 – Fit the model without this variable. Go to Step 3
* Finish – Model is ready
* Bidirectional Elimination :-
* Step 1 - Select a significance level to enter and stay in the model. (e.g SLEnter = 0.05, SLStay = 0.05)
* Step 2 - Perform the next Step of Forward Selection(new variable must have P < SLEnter to enter)
* Step 3 - Perform All Steps of Backword Elimination (old variable must have P < SLSTay to stay)
* Step 4 - Keep going from Step 3 to Step 2 until no new variables can enter and no old can exit
* Finish – model is ready