Linear Regression

```
1.Linear Regression means it is an predective model used to predict the dependent
variables based
on independent variables.
2.Dependent variables are continuous
3.Independents variable may be discrit or may be continuous

linear regression formula
y=mx+c
m>>slop
x>>independent variable
c>>intercept
y>>dependent variable
```

Types of linear regression

```
1.simple LR:
    in this only one dependent variable
 y=mx+c
2.2. Multiple LR: thre is more indepndent variables
    y = m1x1 + m2x2 + ... + mNxN + c
```

Assumptions of Linear Regression

```
there are four assumptions two are before building the model and two after build the
model
1. linearity
2.multicolinearity
3.mormality
4.homoscdasticity
Homoscedasticity of Residual or No Heteroscedacticity:
    Residual should follow Homoscedasctic behaviour
    Fitted values vs Residual
```

1.Linearity

```
1.Linearity means there is linear relationship between dependent and independent
variable
2.for check the linearity we have to  check the coifficent of corroliation which
also called as R value

R value=summetion of (Xi-Xmean)(Yi-Ymean)/rut(Xi-Xmean)^2*(Yi-Ymean)^2

*** Range for R value is -1 to 1 *****
if the R value is >> 0.7 then it is gud prediction
if R<0.3 then it is bad Prediction
```

```
Model accuracy depends on MSE which is Mean Squred Error:

MSE=(Ya-Yp)^2/N
Ya>>y actual dependent variable
Yp>>pridected Dependent variable

if the mse is less then model accuracy is gud
if the mse is high then model acuurcay is bad
```

Gradian Decent Algorithm:

```
1 This is one algorithm to reduce the cost of function

2.this is used for the find out the M and C values

3.this will try the infinite m and c values till we get the best m and c values

4.it uses partial derivative

in this we have to check new m value

if our first m value is 1 and slop is 3 dimentional
in that at m=1 the mse is 100 that is learning step then this algorithm will try new m
values

Mnew= Mold -learning Rate *deriavative MSE /m
learning rate can be=0.01
Mnew= 1-0.01*-1
mnew= 1 +1
mnew= 2

Like this this will try for the multiple m and c values till the we get best M and C
values or Low mse

once we get the the Low MSE at some point this is called as Global MInima
 but when we changing the m and c values after global minima then mse will increase
```

Global Minima

```
1.Global Minima is a point at that point we get the Best m and c values
2.MSE is low
3.when we changing the m and c values after global minima then mse will increase
```

Best Feet Line

```
1.Best feet line is nothin but regresstion line
2.it passes through the multiple points
3.Low MSE
4.Best M and c Values
5.it is trying number of possibilitits for getting best M and c values
```

Evalution Of Linear Regression

```
1.MSE>>mean squred value
2.SSE>>sum of squred error
it is squred difference between y actual and y predicted or we can say squred difference
of residuals

SSE=(ya-yp)^2

3.SSR>>sum of squres due to regression
squred difference between Y predicted and mean of dependent variable

SSR=(Yp-Y mean)^2

4.SST>>sum of squres of toatal error
squred differnce between y actual and mean of dependent variable

SST=(ya-y mean)^2
```

R2 Score

```
1.R2 score or R2 value basically it is a coifficent of Determination
2.it is used for find the gudness of best feet line

R2 score= 1-SSE/SST or (SST-SSE)/SST

R2 score is  1 when SSE is 0
R2 ==1 means all the data points on regression line & it is gud

R2 score is negative when SSE is greater than SST

R2 score is 0 when SSE=SST
and 0 is worst score
```

R2 score in terms of variance

```
R2=var(mean)-var(BFT)/var(mean)

1.there are two terms explained variance and unexplained variance
2.unexplained variance means simply SSE
2.explained variance means differnce between SSE and SST
```

Features of R2 Score

```
there are 4 features of R2 score
R2=0.85
case1 : if the R value is greater than 0.7 then it is gud predictor
```

```
    for that R2 Score will be 0.88 means R2 score is incresing for gud predictors

case2 : if the R value <0.3 means it is bad predictors and R2 score for this will be
0.86
    means this is incresing for bad predictors also

    so this is not correct we dont get gud result

    3.R2 score never decreses
    4.and it is increses for bad predictors this is drawback of R2 Score
```

Adjusted R2 score

```
1. to overcome the drawback of R2 score this is one term Adjusted R2 score
2. this will increses only for bad predictors and decresses for bad predictors

R-2(adjusted R2 score)=(1-R^2)(N-1)/(N-P-1)
N>>number of samples
P>>number of predictors

Adjusted R2 score value always less than R2 value
```

Overfitting and Underfitting

```
1..Overfitting: when the acuuracy on training data is high and accracy on test data is
low
    it is called as overfitting

    means low bias and high variance

2..Underfitting : when the accuracy on train data is low and accuracy on test data is
also low
    then it called as underfitting

    high bias and low variance

3..bias : it dependent on accuracy of train data
    1 if the train data accuracy is high then it is low bias
    2 if the train data accuracy is low then it is high bias

4..vriance :Variance:
    Difference between accuracies of different datasets

high variance : if the differnce between the accuracys of different datasets is more
     1.high accuracy on train data and low accuracy on test data
     2.high accuracy on test data and low accuracy on train data

low variance :  if the differnce between the accuracys of different datasets is less
     1.high accuracy on train data and high accuracy on test data
     2.low accuracy on test data and low accuracy on train data
```

```
Overfitting:
    Train data accuracy  >> High  >> 96%  >> 1000
    Test Data accuracy   >> Low   >> 75%  >> 5000
    Low Bias and High Variance


Underftting:
    Train data accuracy  >> Low >> 70
    Test Data accuracy   >> Low >> 70

    High Bias and Low Variance

Bias >> Accuracy on train data:
    Low Bias >> High Accuracy
    High Bias >> Low Accuracy

High Variance >>
    1. High Train Accuracy and Low Test accuracy  >> More difference
    2. Low Train Accuracy and High Testing Accuracy

    Low Variance:
        1. High Train Accuracy and High Test accuracy
        2. Low Train Accuracy and Low Testing Accuracy
```

Advantages

```
1. Perform exceptionally well on linearly seperable data
2. Easy to implement
3. Overfitting can be reduced by regularization(L1 and L2)
```

Disadvantages

```
1. Linearity
2. Independence
3. Sensitive to outliers
4. Sensitive to missing value
```

Encoding

1.1 One hot encoding

```
**** if tha dataset columns datatype is object then we can use encoding****

1..If we dont know the prefrence of values then we can use one hot encoding
2..suppose in our dataset the columns contains three value gas ,fual,disel but we dont
know the
what is prefrence for gas and ,fual and disel in that case we can use lable encoding
```

```
3..if we using lable encoding means we spliting that original column means dimention
will increase

one hot enconding can be done like

high   after one hotencoding   high  low   medium
low                              1    0     0
medium                           0    0     1
high                             1    0     0
low                              0    1     0
high                             1   0      0

where the value is present it will replace that by one other value will be zero like
this

there is one direct function for ine hot encoding using which isget dummies()
or u can also import libarary and use
one hot encoding function

in get dummies there is one option drop_first u can make it True means it will drop
first
column and will reduce the dimension
```

1.2 Lable Encoding

```
1..If we know prefrence of values then we can use the lable encoding

suppose there are values like
high>>2
low>>0
medium>>1

or
four>>0
five>>1
six>>2

In this case we know the preference or we can give the wattage in this case we can use
the
lable encoding

for lable encoding we can use direct pandas replace function or we can import libarary
```

How we can check the normality

```
all resudials follow the normality curve we can check this
# For check the normality there are four test
1..Density plot
2..shapiro Test
3..normality test
4..ktest
5..QQ plot for visualisation
```

2 Shapiro test /hypothesis

```
there are two hypothesis

1..Null hypothesis: null hypothesis accepted means we follow the null hypo

2.. Alternative hypothesis: if we rejecting the null hypothesis means we following
alternative hypo
i
if the probablity value is greter than 0.05 then we can accept the Null hypothesis
(_p_value=0.05)

if the _p_value>=0.05 means data is normally distributed
```

QQ plot

```
This is for visualisation part but we can not sure on qqplot

if the all points on red line then we can say data is normally distributed but not
surely
```

Homoscadsticity

```
The asmeption of equal variance
```

Outlyers

```
outlyers means means those data points which are far away from the obesrvations
or we can say the those numbers which are out of the range
```

How outlyers are introduced in data

```
1..Data Entry error: we can also call it as a human error
means sometime typing mistake is there
2..measurement or instrument error: suppose we have to measur the bloodpressure we are
mesuring
    that using some instrument but that is not working well in this case we will not
    get correct result this is called as measurement error
3...Intentional Error:Dummy error

4...Sampling Error:Mixing of data from wrong resources
5..Natural Error:Most of the data belongs to this category this is not actually error
```

Impact of outlyers

```
# 1.Reduce the power of stastical anylysis
```

```
2.high impact on mean value and std deviation its shifting towrds the outlyers

3.but there is no impact on meadian if there is any impact then it will be very small or
not too much

4.alogorithm do not perform well in the presence of outlyers (accuray ,mse)
means tere will be impact on accuracy and mse.

5.impact on basic assumption of regression(normality,homoscasticity)
```

How To Detect Outlyers

```
There are some methods to detect the outlyers
1...Z_Score :using this method we can detect the outlyers
    Z_score=(X-Xmean)/std #formula for z_score
X>>element from array
Xmean>>mean of that array
std>>standard deviation of that array

This is equvalent to the standardisation

2...IQR-Method: Inter Quartile Range
we to find quartile q1=np.quantile(array,0.25)
q2=np.quantile(array,0.50) #it is meadian of that column
q3=np.quantile(array,0.75)

IQR=q3-q1

Upper_tail=q3+1.5*iqr
Upper_tail=q1-1.5*iqr
 The values are less than lower_tail and greater than upper_tail that will be the
outlyers
3.Box_plot:this is for visulisation
outlyers is indicated by the  dot in that box.if there is no any dot out of the box
means no outlyers

4..scatter plot
```

```
Handling or Replacing outlyers
1..Delete obervations
for deleting the outlyers first we have to find out the index of that outlyers and then
we can delete
that

2...Imputaions:means we can replace that outlyers by mean meandian,mode or any static
value
    but standard method is replace that by mean or meadian or by mode

3...Transformation:it used to reduce the impact of outlyers
    1.Log Transformation
    2.Normalisation(range is 0 to 1)
    3.Standardization (there is no fix range)
    4.cuberoot transformation
    5.Reciprocal transformation
```

Outlyers impact on algorithm

Algorithms those are snsitive to outlyers

```
1.Linear Regression
2.Logistic Regression
3.K-nearst_Nebougher
4Support vector machine
5.K-means-clustering
```

Algorithms those are not snsitive to outlyers

```
1.Decision Tree
2.Adaboost
3.XGboost
4.random forest
5.naive bays classifier
```