**Linear Regression**

1. Linear Regression means it is a predictive model used to predict the dependent variables based on independent variables.
2. It is used to find a linear relationship between a dependent variable and one or more independent variables.
3. Dependent variables will be continuous.
4. Independent variables may be discrete or may be continuous.

**Linear regression expression**

y=mx+c

m => will be the slope of linear regression

c => will be the intercept y m

x => will be the independent variable y = mx+c

y => will be the dependent variable ya ya yp 

yp

yp yp ya

c ya

x

Linear regression line is also called Best Fit Line.

Regression line will be returning the lowest MSE, which is known as Mean Squared Error.

Linear regression expression helps to get the best M and C values, Best value of M and C will give the accuracy of the model.

**Types of Linear Regression** :

1. Simple Linear Regression:

Simple regression will be having only one dependent variable

y=mx+c

1. Multiple Linear Regression: Multiple Linear Regression is **a statistical technique that uses several explanatory variables to predict the outcome of a response variable**. Multiple regression is an extension of linear regression that uses just one explanatory variable.

y = m1x1 + m2x2 + ... + mNxN + c

**Assumptions of Linear Regression** :

Basically there are four assumptions used to build a model.

1. Linearity 2. Multicollinearity 3.Normality 4.Homoscedasticity

Two assumptions are used before building the model and two assumptions are used after building the model.

Homoscedasticity of Residual or No Heteroscedasticity : Residual should follow Homoscedastic behavior Fitted values vs Residual

**Linearity** :

1. Linearity means there is a linear relationship between dependent and independent variables.

2. For checking the linearity we have to check the coefficient of correlation which is also called the R value.

3. R value=summation of (Xi-Xmean)(Yi-Ymean)/rut(Xi-Xmean)^2\*(Yi-Ymean)^2

Range for R value is -1 to 1 …

if R value is > 0.7 then it is good predictor

if R value is < 0.3 then it is good predictor

Model accuracy will be depends on MSE which is Mean Squared Error :

MSE=(Ya-Yp)^2/N

Ya = Actual data point of dependent variable

Yp = Predicted data point of dependent variable

If the mse is less then will get good accuracy of model

If the mse is high then will get bad accuracy of model

**Gradient Descent Algorithm:**

1. Gradient Descent Algorithm is an algorithm to reduce the cost of function.

2. This is used to 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 slope is 3 dimensional in that at m=1 the mse is 100 that is learning step then this algorithm will try new m values

5. Mnew= Mold -learning Rate \*derivative dMSE /dm learning rate can be=0.01 Mnew= 1-0.01\*-1

mnew= 1 +1 mnew= 2

Like 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

MSE 





step-size

C

m1 m2 m 

Global Minima

1. Global Minima is a point at that point we get the Best m and c values

2. MSE that is mean squared error will be 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 nothing but regression line

2. It passes through the maximum number of data points.

3. At the best fit line MSE will be low.

4. BFL will give best M and c Values.

5. It will trying number of possibilities for getting best M and c values

**Residual :**

1. Residual is nothing but the error between actual points and predicted points.

2. Positive Residual : If data points are above the regression line.

3. Negative Residual : If data points are below the regression line.

**Evaluation Of Linear Regression :**

1. MSE mean squared value.

2. SSE is a sum of squared error, it is squared difference between y actual and y predicted or we can say sum of squared difference of residuals

SSE = (ya - yp)^2

3. SSR is a sum of squares due to regression, difference between Y predicted and mean of dependent variable

SSR= (Yp-Y mean)^2

4. SST sum of squares of total error squared difference between y actual and mean of dependent variable

SST = (ya-y mean)^2

**R2 Score : coefficient of Determination**

1. R2 score or R2 value basically it is a coefficient of Determination

2. It is used for find the goodness of best fit 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 good R2 score

If Negative R2 score then SSE is greater than SST, we cannot improve model if R2 = -1,

If R2 score is 0 when SSE=SST ,means that is worst score

Main drawback of the R2 score is that it sometimes returns a negative value as well.

To dial with such situation we need to use the concept of ‘Adjusted R2 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 difference between SSE and SST

Features of R2 Score :

There are 4 features of R2 score, R2 = 0.85

Case 1 : If the R value is greater than 0.7 then it is good predictor for that R2 Score will be 0.88 means R2 score is increasing for good predictors

Case 2 : If the R value < 0.3 means it is a bad predictor and R2 score for this will be 0.86

Means this is increasing for bad predictors also so this is not correct we don't get good result

3. R2 score never decreases

4. But R2 score increases for bad predictors this is drawback of R2 Score

**Adjusted R2 score :**

1. To overcome the drawback of R2 score need to used Adjusted R2 score

2. This will increases only for good predictors and decreases for bad predictors

R-2(adjusted R2 score)=(1-R^2)(N-1)/(N-P-1)

N is number of samples

P is number of predictors

Adjusted R2 score value always be less than or equal to R2 score

**Overfitting and Underfitting :**

1. Overfitting : When the accuracy on training data is high and accuracy 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. Variance : Difference between accuracies of different datasets

High variance : if the difference between the accuracy 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 difference between the accuracy 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

**Underfitting**:

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 separable data

2. Easy to implement

3. Overfitting can be reduced by regularization(L1 and L2)

**Disadvantage**

1. Linearity

2. Independence

3. Sensitive to outliers

4. Sensitive to missing value

**Encoding**

If the dataset columns data type is object then we can use encoding

Basically there is 4 way to do encoding

1. Label Encoding 2. One Hot Encoding 3. Hash Encoding 4. Binary Encoding

1. If we don’t know the preference of values then we can use **One Hot Encoding**.

2. Suppose in our dataset the columns contain three values gas, fuel, diesel but we don’t know what is preference for gas and fuel and diesel in that case we can use label encoding.

3. If we using label encoding means we splitting that original column means dimension will increase

one hot encoding can be done like

High after one hot encoding 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 one hot encoding using which is get dummies()

We can also import library 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

**Label Encoding :**

1. If we know preference of values then we can use the label 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 label encoding for label encoding we can use direct pandas replace function or we can import library

**Ways to check the normality :**

All residuals follow the normality curve we can check this Need to

from scipy.stats import shapiro, normality, kstest

For check the normality there are four test

1. Density plot

2. Shapiro Test

3. Normality test

4. Kstest

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 probability value is greater 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 qq plot if the all points on red line then we can say data is normally distributed but not surely

**Homoscedasticity**

The assumption of equal variance

**Outliers**

Outliers means those data points which are far away from the observations or we can say the those numbers which are out of the range

**How outliers are introduced in data**

1. Data Entry error: we can also call it as a human error

2. Measurement or instrument error: suppose we have to measure the blood pressure we are measuring 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 outliers**

1. Reduce the power of statical analysis

2. High impact on mean value and std deviation its shifting towards the outliers

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

4. Algorithms do not perform well in the presence of outliers (accuracy ,mse) means there will be impact on accuracy and mse.

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

**Detect Outliers**

There are some methods to detect the outliers

1. Z\_Score 2. IQR 3. Boxplot 4. Scatterplot

1...Z\_Score : Using this method we can detect the outliers

Z\_score = (X-Xmean)/std formula for z\_score

z\_score= np.abs(stats.zscore(array))

X is element from array

Xmean is mean of that array

Std is standard deviation of that array

This is equivalent to the standardisation

2. IQR-Method: InterQuartile Range

We need to find quartile

q1=np.quantile(array,0.25)

q2=np.quantile(array,0.50) #it is median of that column

q3=np.quantile(array,0.75)

IQR = q3-q1 Lower\_tail = q1-1.5\*iqr Upper\_tail = q3+1.5\*iqr

The values are less than lower\_tail and greater than upper\_tail that will be the outliers

3. Boxplot : This is for visualization, outliers are indicated by the dot in that box.

If there is no any dot out of the box means no outliers

4. scatterplot

Handling or Replacing outliers

1. Delete observations

for deleting the outliers first we have to find out the index of that outliers and then we can delete that

2. Imputations : means we can replace that outliers by mean median, mode or any static value

but standard method is replace that by mean or median or by mode

3. Transformation : It used to reduce the impact of outliers

1. Log Transformation

2. Normalisation(range is 0 to 1)

3. Standardization (there is nfix range)

4. cube root transformation

5.Reciprocal transformation

**Outliers impact on algorithm :**

**Algorithms those are sensitive to outliers**

1. Linear Regression

2. Logistic Regression

3. K-nearest Neighbor

4. Support vector machine

5. K-means-clustering

**Algorithms those are not sensitive to outliers :**

1. Decision Tree

2. Adaboost

3. XGboost

4. Random forest

5. Naive bays classifier

**Transformation**

1. scaling:

1. Normalization

2. Standardization

2. Log Transformation

Observation will not deleted

**Normalization**

Range >> 0 to 1

Xnorm = (X - Xmin)/(Xmax-Xmin)

Need to import libraries

from sklearn.preprocessing import MinMaxScaler, StandardScaler