Machine Learning:

Machine learning enavles a machine to automatically learn from data, improve performance from experience and predict things without being explicity programmed.

TRADITIONAL LEARNING -- (input + logic = output)

machine learning --> (input & outup) == +

mcahine larning has 2 models

Training Phase and Testing Phase

- Historical data is known as Training Data.

- Machine learning is a combination of Computer science and statistical data.

-Input past data (training)-- Machine learning algorithm(learn from data)--Building logical models--(new data)---output

Machine learning follows the above process

If we add the data in the middle also it performs as per the new data

Machine Learning Features:

-Machine learning uses data to detect various patterns in a given dataset.

- It can learn from past data and improve automatically.

-It is a data driven technology (it takes data automatically)

-Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Classifications of ML:

1. Supervised Learning: Supervised learning is a type of machine learning method, in which we provide a sample labeled data to the machine learning system in order to train it and on that basis it predicts the output.
- It tries to generate the input data to output data
- spam filtering
- CLassifed into 2 categories
classification

regression

2.Unsupervised Learning: The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns - machine learns without any supervision

- We cant predict the output

-CLassifed into 2 categories

Clustering

Association

3.Reinforcement learning: Feedback based learning method.

learning agent gets a reward for each right action

gets penalty for each wrong action.

The agent learns automatically with these feedbacks and improves its performance

Ma chine Learning life cycle:

1.Gathering Data:

2.Data preparation:

Raw data-structure data-Data preprocessing-EDA-Insights,Reports,Visual Graphs

3.Data wrangling:Issues--Missing values,Duplicate data, Invalid data

4.Analyse data:

Selection of analytical techniques

Building models

Review the result

5.Train model:

6.Test Model:

7.Deployment

Linear Regression

- It makes predictions for continuous/real or numeric variables such as salaes,salary,age,product price etc.,

- Linear regression algrmthm shows a linear relationship between a dependent(y) and one or more independent variables(x)

Types of linear regression

1.SImple Linear Regression

If a single independent variable is used to predict the value of numeric dependent variable.

2.Multiple linear regression:

If more than one independent variable is used to predict the value of numerical dependent variable.

Linear regression Line: A linear line shows the relationship between the dependent and independent variables is called a regression line.

Positive Linear relationship:

If the dependent variable increases on the Y-axis and independent variable increases on X-axis.

Negative Linear relationship:

If the dependent variable decreases on the Y-axis and independent variable increase on the X-axis.

Best Fit line:

- It is the error between predicted values and actual values should be minimized.

- The best fit line will have the least error

- to find the best fit line, so to calculate this we use cost function

Cost Function:

The cost function is used to find the accuracy of the mapping function.

Which maps the input variable to output variable.

MSE-Mean squared error average of squared error occurred between the predicted values and actual values.

Residuals:

- The distance between the actual value and predicted value is called residual.

- If the observed points are far from the regression line, then the residual will be high and so cost function will high

-If scatter points are close to regression line, then the residual will be small and hence the cost function.

Gradient Descent:

Gradient descent is used to minimize the MSE by calculating the gradient off the cost function.

A regression model uses gradient descent to upgrade the coefficients of the line by reducing the cost function.

It is done by random selecting values of coefficient

They iteratively update the values to reach the minimum cost function.

R-Squared:

R-Squared is a statistical method that determines the goodness of fit

It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%

The high value of R-squared determines the less difference between the predicted values and actual values and hence represents a good model

It also called a coefficient of determination.

R-squared=Explained variation/Total variation

multicollinearity :

it means high correlation between the independent variables.

Due to multicollinearity it may difficult to find the true relationship between the predictors and target variables.

Simple Linear Regression:

Simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response/Dependent variable (Y).

It is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable.

- Simple linear Regression is that the dependent variable must be a continuous value.

- The independent variable can be measured on continuous or categorical values.

Model the relationship between the two variables:

- Such as the relationship between Income and expenditure, experience and Salary etc.,

Forecasting New Observations:

- Such as weather forecating according to temparature,

- revenue of the company according to the investiment in a year, etc.,

y=a0+a1+epsilon(E)

a0=Intercept of the Regression Line

a1= Slope of the regression line

epsilon(E)- The error term

Multiple Linear Regression:

- Multiple Linear Regression is one of the important regression algorithms which models the

linear relationship between a single dependent continuous variable and more than one independent variable.

- Key points:

For MLR, the dependent variable(Y) must be continuous

But the independent variable may be continuous or categorical form.

- Each feature variable must model the linear relationship with the dependent variable.

-MLR tries to fit a regression line through a multidimensional space of data points.

In MLR, the dependent variable(Y) is a linear combination of multiple predictor variables x1,x2,x3....xn

It is a enhancement of Simple Linear Regression.

- A linear relationship should exist between the Dependent and predictor variables.

- The regression residuals must be normally distributed.

- MLR assumes little or no multicollinearity(correlation between the independent variable) in data

- Main steps of deploying the MLR Model:

Data Pre-Processing steps

Fitting the MLR model to the training set

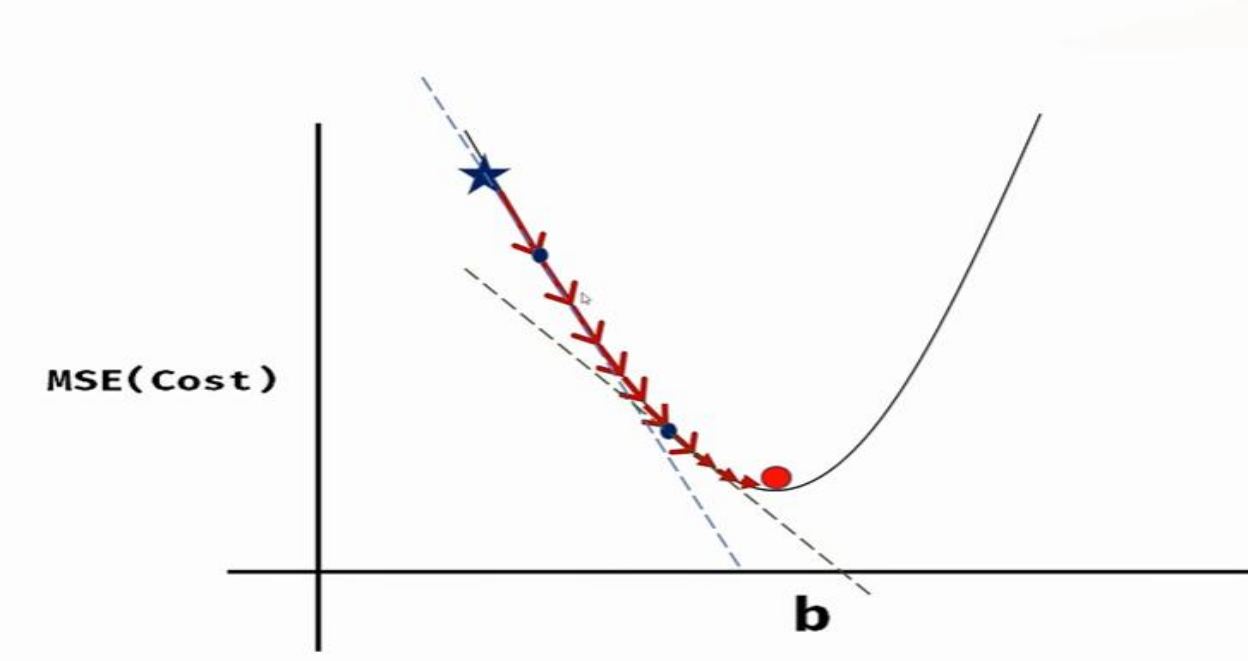
Predicting the result of the test set

==========

Gradient Descent:

It is an algorithm that finds best fit line for given training dataset.

We perform more steps or iterations to get a best fit line. Until reach the local minima



Backward Elimination:

steps of Backward Elimination

step1:- Firstly need to select a significance level to stay in the model.(SL=0.05) (alpha)(hypothesis)

step2:- Fit the complete model with all possible predictors/independent variables.

step3:- Choose the predictor which has the highest P-Value, such that

a. If p-value>SL go to step 4

b. Else finish, and our model is ready

step 4: Remove the predictor.

steps5: Rebuild and fit the model with the remaining variables.

Need of Backward elimination:

- An optimal Multiple Linear Regression Model

Normalization and Standardization:
Numpy: for numerical operations
Pandas: for data manipulation and analysis
Matplotlib.pyplot: for plotting graphs
SimpleImputer: is used to handle missing values
Fit: learns the imputer model from the data (calculates the mean in this case)
Transform: applies the imputer to fill the missing values in 'x'
LabelEncoder: converts categorical data to numerical data

Fit_transform: encodes the categorical data and replaces it with numerical values.

Feature Scaling:

Unit

Magnitude :25 yrs

Normalization: Normalization helps you to scale down the feature between 0-1

Standardization: It also known as z-score normalization transforms the data to have a mean of 0 and sd of 1

Standardization which helps you to scale down your feature based on standard normal distribution(bell curve)

Over there usually mean=0, sd=1

Random Forest Regressor Parameter:

"squared_error" for the mean squared error, which is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node,

"friedman_mse", which uses mean squared error with Friedman's improvement score for potential splits,

"absolute_error" for the mean absolute error, which minimizes the L1 loss using the median of each terminal node,

"poisson" which uses reduction in Poisson deviance to find splits. Training using **"absolute_error"** is significantly slower than when using "squared error".

Polynomial Regression:

Polynomial regression is nothing but it maintains a relationship between Dependent and independent variable till nth degree polynomial

$Y=mx+c$

$Y = a + b_1X_1 + b_2X_2$, becomes the quadratic polynomial

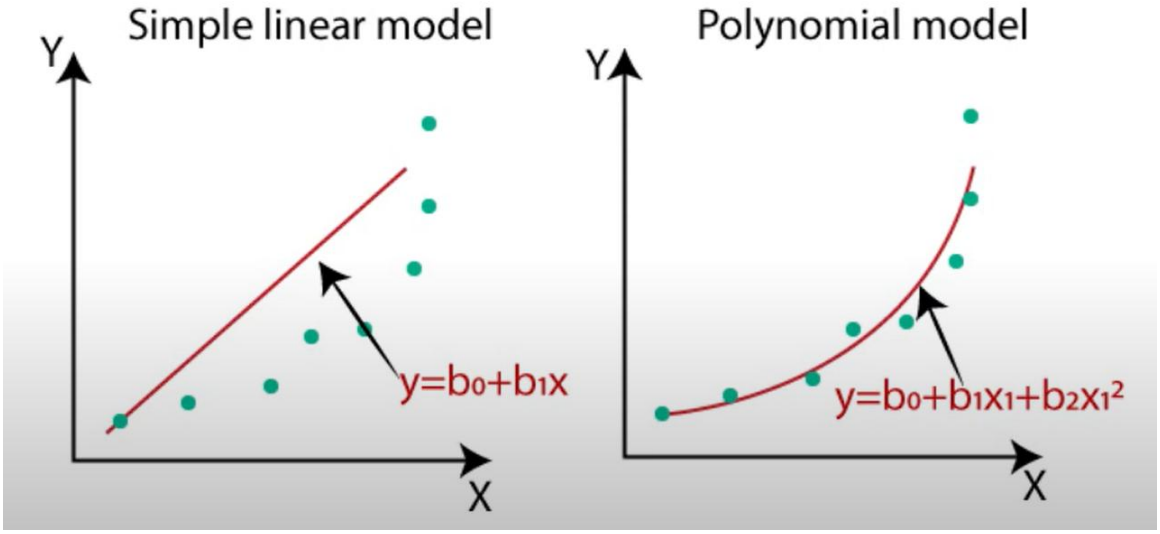
$Y = a + b_1X + b_2X^2$.

It is also a special case of a multi linear regression

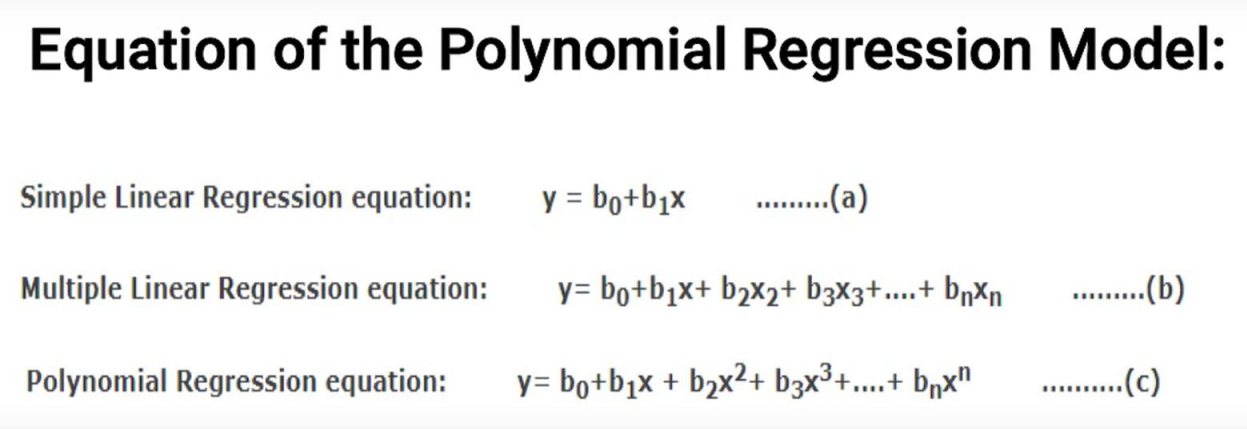
Purpose: Polynomial regression is used when the data shows a non-linear relationship that cannot be captured by a simple linear model.

Linear data: Data That is organized remain sequential order which each elements connected to previous and next elements example of linear data encrypts arrays lists pack and queues.
None Linear Data: Data that is not organized in a sequential order and can have multiple relationship with other elements example of non linear data include tree graphs and hash tables.
Dataset should be non linear y because we will train the data from non linear to linear
Here we will use linear regression model to reduce the complexity.(non-linear and dataset)

Need of polynomial:
If we apply a linear model to a non-linear dataset, the output can be drastically inaccurate. The loss function measures the error rate, and when the error rate increases, the accuracy decreases significantly.



Equation of Polynomial Regression Model:



Steps for polynomial regression

- Data Pre-processing
- Build a linear regression model and fit it to the data set
- Build a polynomial regression model and fitted to the dataset
- Visualize the result for linear regression and polynomial regression model
- Predict the output

Support Vector Regressor: SVR is a type of support vector mission that is used for regression task unlike typical regression models that aim to minimize the error between predicted and actual values

K-Nearest Neighbors (KNN):

It finds the 'k' closest data points neighbors to a given point.

Classification: For classification it assigns the most common label among these neighbors to the point.

Regression: For regression it averages the value of the neighbors to predict the value for the point.

Decision Tree: Imagine you have a complex decision to make, like figuring out what to wear based on the weather and the occasion. A decision tree helps you make this decision by breaking it down into simpler, step-by-step choices, much like a flowchart.

Random Forest: Imagine you want to predict something, like the weather or what movie you might like. Instead of asking just one friend (a single decision tree), you ask a whole bunch of friends (a forest of decision trees), and each gives their opinion. Then, you take the most common answer (for classification) or the average (for regression).

Classification

Binary classification, Multi class classificatin

Lazy Learners: They store the training data and wait until a query is made, and then generalize the data for each query to make a prediction. This means they have a **slow prediction** phase because they perform computations during this phase. (KNN)

Eager Learners: Develop the classification model based on the training dataset, Here it will developed before the getting the test data. It is opposite to the lazy learners

This results in a **slow training** phase but **quick predictions** once the model is built. (Decision Tree, Neural Network, Support Vector machine)

Linear Models:

Logistic Regression

Support Vector Machine

Non-linear models:

KNN,Kernel-SVM, NAÏVE BAYES, DECISION TREE CLASSIFICATION, RANDOM FOREST CLASSIFICATION.

Evaluating a Classification model:

1. **Log Loss or Cross-Entropy Loss**
   Used for evaluate classifier performance (0-1)
   0- Good binary classification
   1- Not Good binary classification.
2. **Confusion Matrix**
   It describes the performance of the model . It also known as error matrix

	Actual Positive	Actual Negative
Predict positive	TP	FP
Predict Negative	FN	TN

3. **AUC-ROC curve**
   Used for multi class classification
   Based on TPR and FPR we build a Multi class classification model
   TPR- True Positive Rate
   FPR- False Positive Rate
   AUC: Area Under the Curve
   ROC: Receiver Operating Characteristics Curve

Use cases of Classification Algorithms

Email spam detection, Speech Recognition, Identifications of Cancer tumor cells, Drug classification Biometric Identification etc.,

Classification:

Linear Regression	**Classification**
1. Home prices 2. Weather 3. Stock price	1. Email is spam or not 2. Will customer buy life insurance 3. Will customer buy a vehicle 4. Which party a person is going to vote for? 1.BJP 2.BRS 3.CONGRESS 4.MIM (Multi class classification) Yes/No—Binary classification
Predicted value is continuous	Predicted value Is categorical

Logistic regression is one of the techniques used for classification

when name itself mention logistic regression why we call this as logit call as classification when the data has outlier misclassification problem happens thats why introduce to probability function that probability function - sigmoid function the data has outlier or no outler it does not impact much sigmoid probability function 1 / 1 + E-y

LINEAR REGRESSION - BEST FIT LINE

LOGISTIC REGRESSION - BEST FIT CURVE

why logistic regression is classification algorithm ?

When the data has outlier sigmoid adjust those outliers deep learning sigmoid activation function range of sigmoid 0-1.

Sigmoid Function: is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range 0 and 1

The value of the logistic regression must between 0-1 , it forms a curve like 'S' form. The S-form curve is called the Sigmoid function or the logistic regression.

Assumptions for Logistic Regression:

The dependent variable must be categorical in nature.

The independent variable should not have multi collinearity

Multicollinearity (it means high correlation between the independent variables.**)**

Steps in Logistic Regression:

Data preprocessing step

Import libraries, load the data, extract independent and dependent variable, splitting the data into training and testing set, Feature scaling

Fitting Logistic Regression to the Training set:

.fit(x_train,y_train)

Predict the test result

Predict the the y_pred with x_test

Test accuracy of the result (Creation of Confusion matrix)

- **We fit the model with (x_train,y_train)**
- **We predict the model with Y-pred=dataset.predict(x_test)**
- **Build a confusion matrix with(y_test,y_pred)**
- **Accuracy (y_test,y_pred)**
- **Train score(bias): (x_train, y_train)**
- **Test_score(variance): (x_test,y_test)**

Visualizing the test set result

We have three phases

Training phase

Testing phase

Validation phase (Future prediction)

The given data we split it into training and testing phase. The range of training and testing is 70|30 or 80|20 or 85|25

CONTROL FLOW:

1. Business Understanding
2. Attribute Understanding (Feature Understanding)
3. Data Cleaning (DATE-MM:DD:YY, TIME-HH:MM:SS) REGEX-CLEAN THEM ALL
4. Independent and Dependent variable
5. Apply statistical graph, EDA, Correlation Graph, Outlier.
6. CHOOSE THE MODEL BASED ON DEPENDENT VARIABLE
7. We have to analyze the dataset is Regression, Classification, Clustering model .
8. Import Libraries
9. Split the data X&Y
10. x_train,x_test,y_train,y_test
11. Feature Scaling (Standard Scaler)
12. Call the algorithm
13. Fit the model (x_train,y_train)
14. Build y_pred by pass x_test
15. Compare the model to y_test vs y_pred
16. Build the confusion matrix
17. Model is Overfitting or Underfitting
18. Once you got the best fit model
19. Every model you have to validate test
20. You have to all rest of required algorithm with validation
21. Best accuracy model will go for pickle
22. Website create
23. ML model display in the website
24. Entry (Records are added in Backend)
25. Predict the forecasting
26. Business Understand Future Market
27. Retrain the model with new data
28. Continue the process