DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

III BTECH II SEMESTER 2021-22

DATA ANALYTICS UNIT-III

Regression Concepts

- It is a Predictive modelling technique where the target variable to be estimated is continuous.

Examples of applications of regression

- Applications of regression are numerous and occur in almost every filed, including engineering, the physical and the social sciences, and the biological sciences.
- Predicting a stock market index using other economic indicators.
- projecting the total sales of a company based on the amount spent for advertising

Regression

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').

The terminology you will often listen related to regression analysis is:

- **Dependent variable or target variable:** Variable to predict.
- **Independent variable or predictor variable:** Variables to estimate the dependent variable.
- **Outlier:** Observation that differs significantly from other observations. It should be avoided since it may hamper the result.
- **Multicollinearity:** Situation in which two or more independent variables are highly linearly related.

Regression is the task of learning a target function 'f' that maps each attribute set x into a continuous-valued output y.

For an input x, if the output is continuous, this is called a **regression problem**. For example, based on historical information of demand for smart phone in our mobile shop, you are asked to predict the demand for the next month. Regression is concerned with the prediction of continuous quantities.

Why do we use Regression Analysis?

As mentioned above, regression analysis estimates the relationship between two or more variables. Let's understand this with an easy example:

Let's say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the

growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

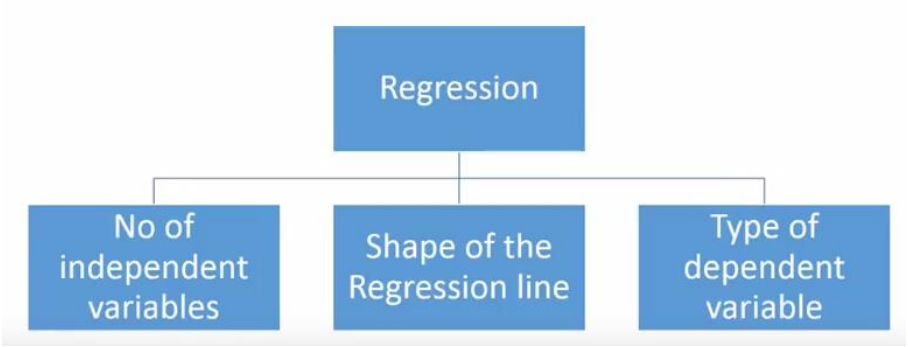
There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

How many types of regression techniques do we have?

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).



For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven't used before. But before you start that, let us understand the most commonly used regressions:

The goal of regression

- To find a target function that can fit the input data with minimum error.
- The error function for a regression task can be expressed in terms of the sum of absolute or squared error:

$$\text{Absolute Error} = \sum_i |y_i - f(\mathbf{x}_i)|$$

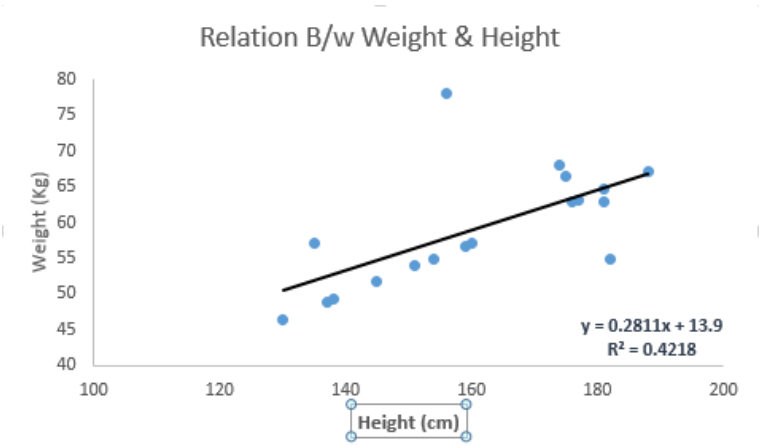
$$\text{Squared Error} = \sum_i (y_i - f(\mathbf{x}_i))^2$$

Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

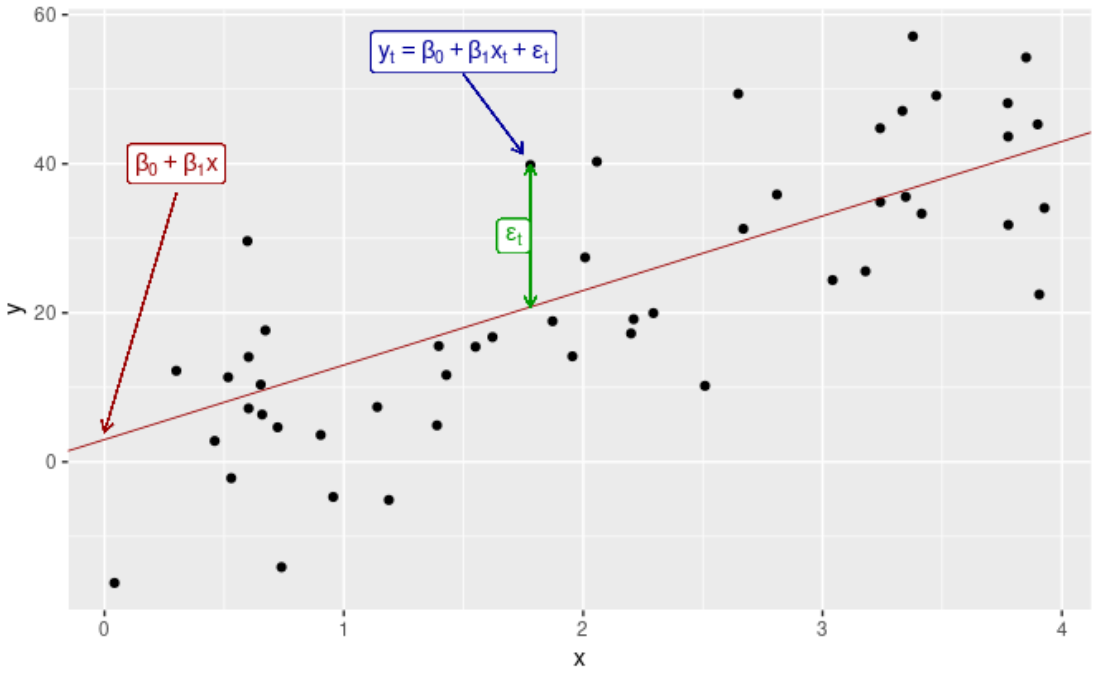


The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is “How do we obtain best fit line?”.

In the simplest case, the regression model allows for a linear relationship between the forecast variable y and a single predictor variable x:

$$y_t=\beta_0+\beta_1 x_t+\varepsilon_t.$$

An artificial example of data from such a model is shown in Figure. The coefficients β_0 and β_1 denote the intercept and the slope of the line respectively. The intercept β0 represents the predicted value of y when x=0. The slope β_1 represents the average predicted change in y resulting from a one unit increase in x.



The simplest case of linear regression is to find a relationship using a linear model (i.e line) between an input independent variable (input single feature) and an output dependent variable. This is called **Bivariate Linear Regression.**

On the other hand, when there is a linear model representing the relationship between a dependent output and multiple independent input variables is called **Multivariate Linear Regression**.

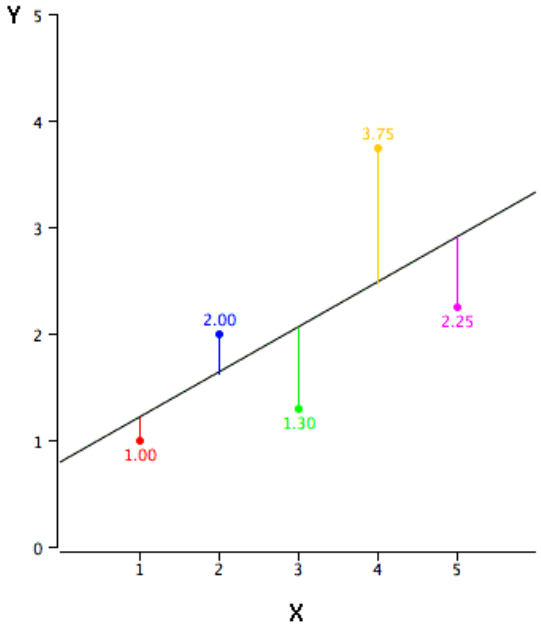
The dependent variable is continuous and independent variables may or may not be continuous. We find the relationship between them with the help of the best fit line which is also known as the **Regression line.**

How to obtain best fit line (Value of a and b)?

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

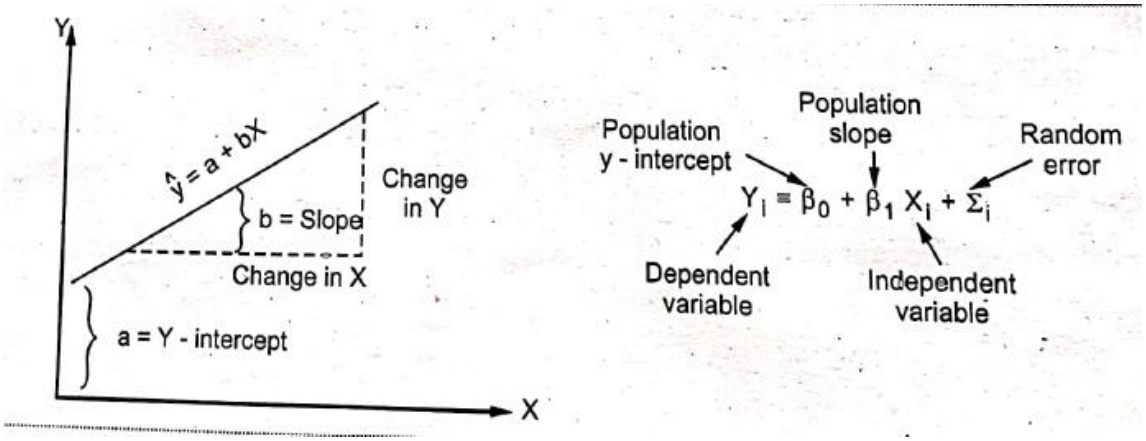
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

$$\min_{w} ||Xw - y||_2^2$$



- We can evaluate the model performance using the metric **R-square**. In multiple linear regressions, multiple equations are added together but the parameters are still linear.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$



Important Points:

- There must be **linear relationship** between independent and dependent variables
- Multiple regressions suffer from **multicollinearity, autocorrelation, heteroskedasticity**.
- Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
- In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.

Blue Property Assumptions

The least squares modeling procedure is the best linear unbiased estimator .i.e., BLUE (Best Linear Unbiased Estimator). The simple model needs five fundamental assumptions to be satisfied and the multiple regression model needs six assumptions to be satisfied. Among this four assumptions are related to model' residuals. They are as follows.

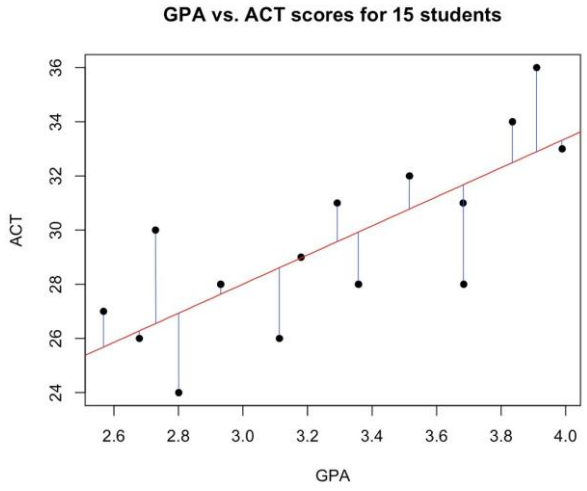
- ✓ The residuals are distributed normally with zero mean E(μ)=0
- ✓ The residuals maintain constant variance (square)
- ✓ The successive residuals are not correlated and there is no chance for auto correlation.
- ✓ The X variables are non-stochastic. They are not correlated with residuals.

Here, the first assumption of zero mean is fulfilled due to the nature of least square estimation. The assumption of normal distribution of residuals is not concerned until the BLUE property is concerned. The Gauss-Markov theorem needs the residuals to maintain zero mean and constant variance. The hypothesis testing however needs normality of residuals. The remaining three assumptions are important in DLS estimation. They are not held always. The performance of forecasting gets affected when any of the three assumptions is violated.

Least Square Estimation or Least Square Method

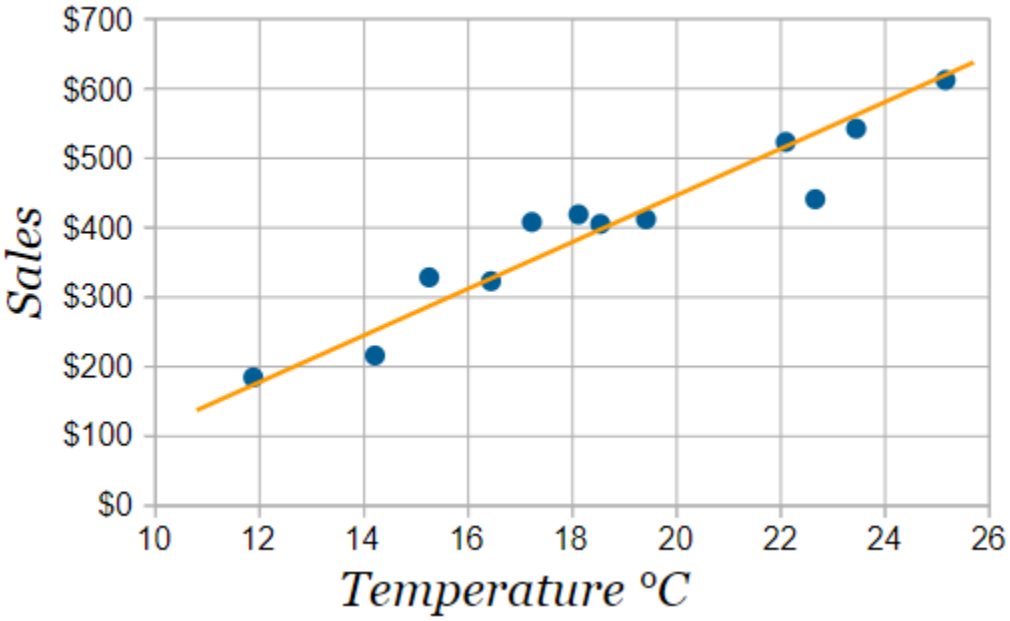
- The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve. Least squares regression is used to predict the behavior of dependent variables.
- Ordinary least squares is a method used by linear regression to get parameter estimates.
- This entails fitting a line so that the sum of the squared distance from each point to the regression line (residual) is minimized.

- Let's visualize this in the diagram below where the red line is the regression line and the blue lines are the residuals.



Line of Best Fit

Imagine you have some points, and want to have a **line** that best fits them like this:



We can place the line "by eye": try to have the line as close as possible to all points, and a similar number of points above and below the line.

But for better accuracy let's see how to calculate the line using **Least Squares Regression**.

The Line

Our aim is to calculate the values **m** (slope) and **b** (y-intercept) in the equation of a line

$$y = mx + b$$

Where:

- **y** = how far up
- **x** = how far along
- **m** = Slope or Gradient (how steep the line is)
- **b** = the Y Intercept (where the line crosses the Y axis)

Steps

To find the line of best fit for **N** points:

Step 1: For each (x,y) point calculate x^2 and xy

Step 2: Sum all x, y, x^2 and xy, which gives us Σx, Σy, Σx^2 and Σxy (Σ means "sum up")

Step 3: Calculate Slope **m**:

$$m = \frac{N \Sigma(xy) - \Sigma x \Sigma y}{N \Sigma(x^2) - (\Sigma x)^2}$$

(N is the number of points.)

Step 4: Calculate Intercept **b**:

$$b = \frac{\Sigma y - m \Sigma x}{N}$$

Step 5: Assemble the equation of a line

$$y = mx + b$$

Done!

Example

Let's have an example to see how to do it!

Example: Sam found how many **hours of sunshine** vs how many **ice creams** were sold at the shop from Monday to Friday:

"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

Let us find the best **m** (slope) and **b** (y-intercept) that suits that data

y = mx + b

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Step 1: For each (x,y) calculate x^2 and xy:

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 2: Sum x, y, x^2 and xy (gives us Σx, Σy, $Σx^2$ and Σxy):

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	**Σy: 41**	**$Σx^2$: 168**	**Σxy: 263**

Also **N** (number of data values) = **5**

Step 3: Calculate Slope **m**:

$$\mathbf{m} = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

$$= \frac{5 \times 263 - 26 \times 41}{5 \times 168 - 26^2}$$

$$= \frac{1315 - 1066}{840 - 676}$$

$$= \frac{249}{164} = 1.5183\ldots$$

Step 4: Calculate Intercept **b**:

$$b = \frac{\Sigma y - m\,\Sigma x}{N}$$

$$= \frac{41 - 1.5183 \times 26}{5}$$

$$= 0.3049...$$

Step 5: Assemble the equation of a line:

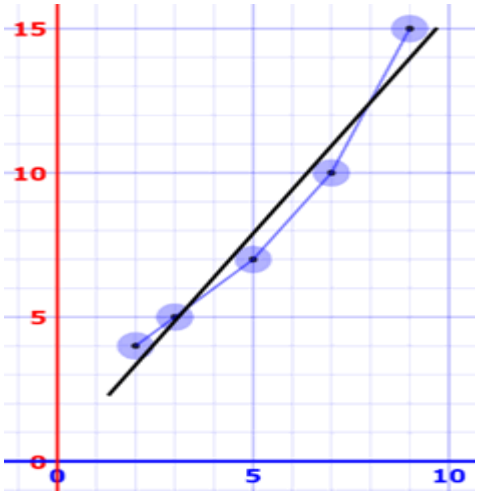
$$y = mx + b$$

$$y = 1.518x + 0.305$$

Let's see how it works out:

x	y	y = 1.518x + 0.305	error
2	4	3.34	−0.66
3	5	4.86	−0.14
5	7	7.89	0.89
7	10	10.93	0.93
9	15	13.97	−1.03

Here are the (x,y) points and the line **y = 1.518x + 0.305** on a graph:



Sam hears the weather forecast which says "we expect 8 hours of sun tomorrow", so he uses the above equation to estimate that he will sell

$$y = 1.518 \times 8 + 0.305 = 12.45 \text{ Ice Creams}$$

Sam makes fresh waffle cone mixture for 14 ice creams just in case.

And for Multiple Linear regression since we have more than 2 independent variables the equation becomes:

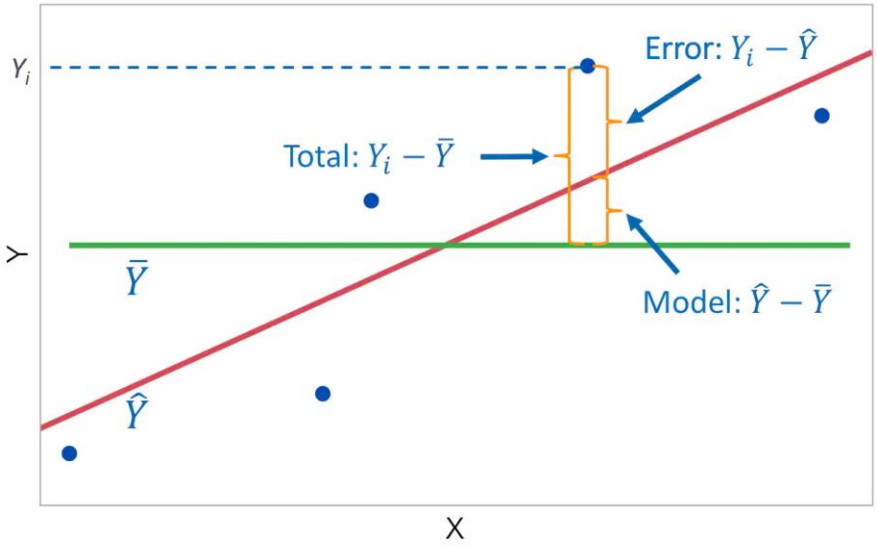
$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots\ldots + x_n\beta_n$$

Where $\boldsymbol{\beta_0}$ is the Y-intercept of the regression line

$\boldsymbol{\beta_1}$ *Is the slope of the regression line*

$\mathbf{X_i}$ *Is the explanatory variable*

Now the question that comes into mind is, what error is this? Can we visualize it? How do we find it? In a linear model or any model we don't have to worry about the mathematical part, everything is done by the model itself.



Let's interpret the graph above. In linear regression the best fit line will be somewhat like this, the only difference will be the number of data points. To make it easier I have taken a fewer number of data points.

Suppose there's a variable Yi, The distance between this Yi and the predicted value is what we call **"SUM OF SQUARED ESTIMATE OF ERRORS" (SSE) .** This is the unexplained variance and we have to minimize it to get the best accuracy.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The distance between the predicted value y_hat and the mean of the dependent variable is called **"SUM OF SQUARED RESIDUALS" (SSR).** This is the explained variance of our model and we want to maximize it.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

The total variation in the model **(SSR+SSE=SST)** is called **"SUM OF SQUARED TOTAL" .**

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

What kind of relationship can a Linear Regression show?

1. **Positive Relationship** – When the regression line between the two variables moves in the same direction with an upward slope then the variables are said to be in a Positive Relationship, it means that if we increase the value of x (independent variable) then we will see an increase in our dependent variable.

2. **Negative Relationship** – When the regression line between the two variables moves in the same direction with a downward slope then the variables are said to be in a Negative Relationship it means that if we increase the value of an independent variable (x) then we will see a decrease in our dependent variable (y)

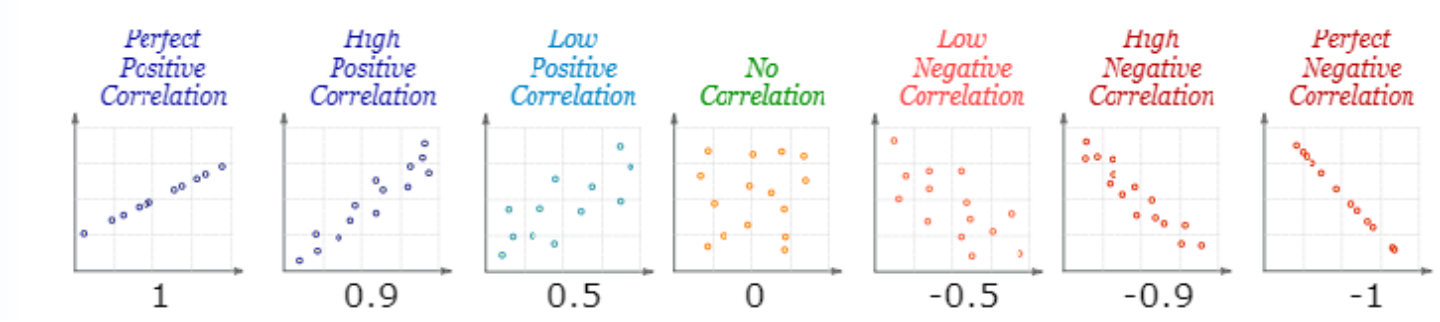
3**. No Relationship** – If the best fit line is flat (not sloped) then we can say that there is no relationship among the variables. It means there will be no change in our dependent variable (y) by increasing or decreasing our independent variable (x) value.

Correlation

- When two sets of data are strongly linked together we say they have a **High Correlation**.
- The word Correlation is made of **Co-** (meaning "together"), and **Relation**

Note:

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases
- A correlation is assumed to be **linear** (following a line).



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

The value shows **how good the correlation is** (not how steep the line is), and if it is positive or negative.

Variable Rationalization

- It is process of clustering data sets into more manageable parts for optimizing the query performance.
- It is used to divide the data but in different way. This process can be assumed as grouping of objects by attributes.
- It is method that increases the performance of bigdata operations.
- Variable rationalization is different from partitioning where every partition contains segments of files.

Advantages & Disadvantages

- It generates faster responses for queries such as partitioning.
- Joins at map side are quicker because of equal volumes of data in every partition.
- Improved performance.
- Provides tools to improve the performance of big data operations.

Disadvantages

- Programmers need to manually load equal amounts of data.
- Programmers need to understand the data before applying the tools.

Model Building

Model building is process of setting various methods to collect data, understanding and then focusing on data. The importance of data must be known to find a statistical or a simulation model to gain understanding and to even make predictions. All these things are important, model building is an important skill to obtain in every field of science.

This process is very much true to scientific method by making the learn things through models to be useful to gain understanding of investigated things and to make the predictions which are true for testing. The process of building variable models involves asking of queries, gathering and manipulating data, building of models and even ultimately testing and evaluating them.

We are going to discuss life cycle phases of data analytics in which we will cover various life cycle phases and will discuss them one by one.

Data Analytics Lifecycle:

The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Phase 1: Discovery –

- ✓ The data science team learns and investigates the problem.
- ✓ Develop context and understanding.
- ✓ Come to know about data sources needed and available for the project.
- ✓ The team formulates initial hypothesis that can be later tested with data.

Phase 2: Data Preparation –

- ✓ Steps to explore, preprocess, and condition data prior to modeling and analysis.
- ✓ It requires the presence of an analytic sandbox; the team executes, load, and transform, to get data into the sandbox.
- ✓ Data preparation tasks are likely to be performed multiple times and not in predefined order.
- ✓ Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning –

- ✓ Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- ✓ In this phase, data science team develops data sets for training, testing, and production purposes.
- ✓ Team builds and executes models based on the work done in the model planning phase.
- ✓ Several tools commonly used for this phase are – Matlab, STASTICA.

Phase 4: Model Building –

- ✓ Team develops datasets for testing, training, and production purposes.
- ✓ Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- ✓ Free or open-source tools – Rand PL/R, Octave, WEKA.
- ✓ Commercial tools – Matlab , STASTICA.

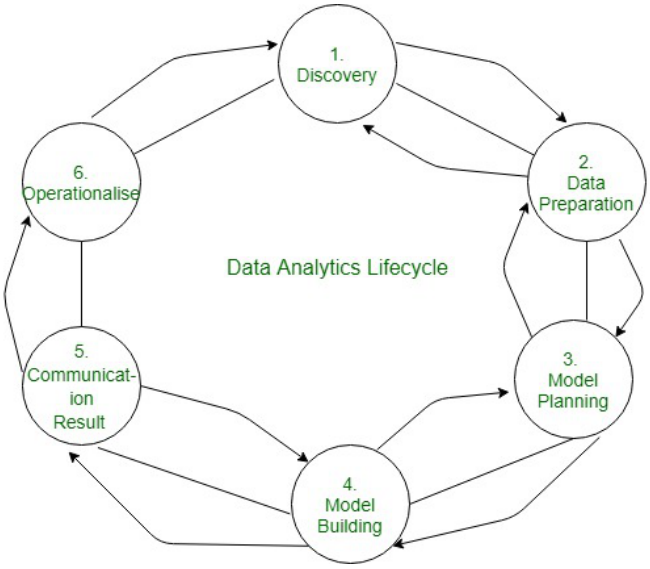
Phase 5: Communication Results –

- ✓ After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- ✓ Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- ✓ Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

Phase 6: Operationalize –

- ✓ The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- ✓ This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- ✓ The team delivers final reports, briefings, codes.

- ✓ Free or open source tools – Octave, WEKA, SQL, MADlib.



Logistic Regression

Classification techniques are an essential part of machine learning and data mining applications. Approximately 70% of problems in Data Science are classification problems. There are lots of classification problems that are available, but the logistic regression is common and is a useful regression method for solving the binary classification problem. Another category of classification is Multinomial classification, which handles the issues where multiple classes are present in the target variable.

Logistic Regression can be used for various classification problems such as spam detection. Diabetes prediction, if a given customer will purchase a particular product or will they churn another competitor, whether the user will click on a given advertisement link or not, and many more examples are in the bucket.

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem. Its basic fundamental concepts are also constructive in deep learning. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables.

Definition of Logistic Regression

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

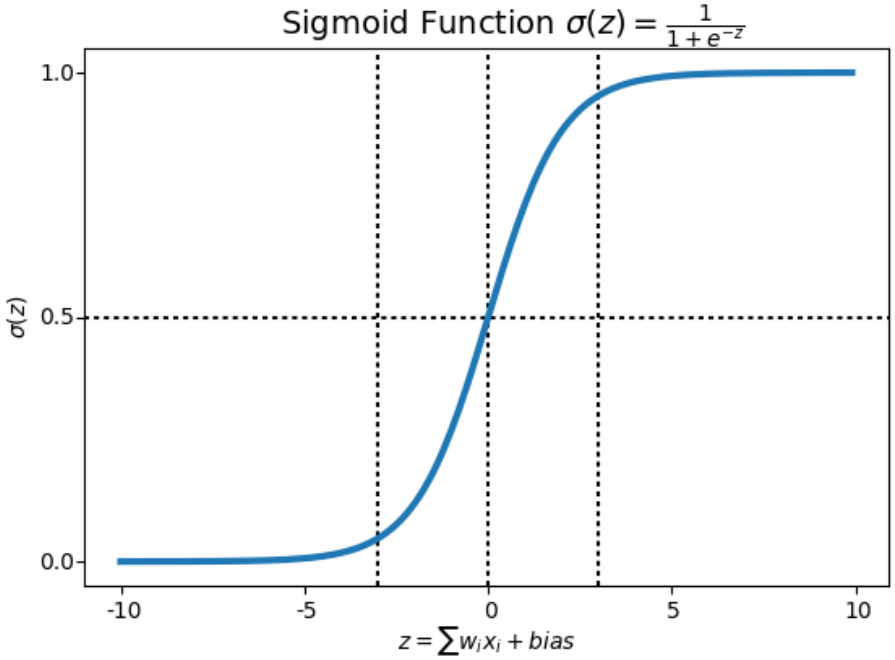
We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

What is the Sigmoid Function?

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



Sigmoid Function Graph

$$f(x) = \frac{1}{1+e^{-(x)}}$$

The sigmoid function, also called **logistic function** gives an 'S' shaped curve that can take any real-valued number and map it into a value between 0 and 1. If the curve goes to positive infinity, y predicted will become 1, and if the curve goes to negative infinity, y predicted will become 0. If the output of the

sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it as 0 or NO. The output cannot For example: If the output is 0.75, we can say in terms of probability as: There is a 75 percent chance that patient will suffer from cancer.

Types of Logistic Regression

- **Binary Logistic Regression**: The target variable has only two possible outcomes such as Spam or Not Spam, Cancer or No Cancer.

  Example:
  - ✓ Whether or not to lend to a bank customer (outcomes are yes or no).
  - ✓ Assessing cancer risk (outcomes are high or low).
  - ✓ Will a team win tomorrow's game (outcomes are yes or no).

- **Multinomial Logistic Regression**: In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. For example, these variables may represent "Type A" or "Type B" or "Type C".

  Example:
  - ✓ Color(Red,Blue, Green)
  - ✓ School Subjects (Science, Math and Art)

- **Ordinal Logistic Regression**: In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent "poor" or "good", "very good", "Excellent" and each category can have the scores like 0,1,2,3.

  Example:

  Medical Condition (Critical, Serious, Stable, Good)
  Survey Results (Disagree, Neutral and Agree)

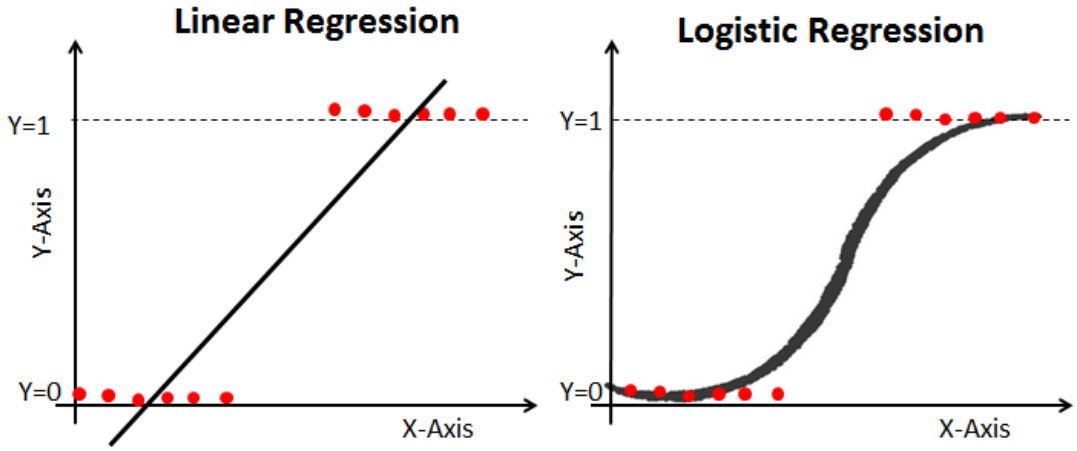
Properties of Logistic Regression:

- The dependent variable in logistic regression follows Bernoulli Distribution.
- Estimation is done through maximum likelihood.
- No R Square, Model fitness is calculated through Concordance, KS-Statistics.

Linear Regression Vs. Logistic Regression

Linear regression gives you a continuous output, but logistic regression provides a constant output. An example of the continuous output is house price and stock price. Example's of the discrete output is predicting whether a patient has cancer or not, predicting whether the customer will churn. Linear

regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.



Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable.

What Is Logistic Regression Used For?

Here is a more realistic and detailed scenario for when logistic regression might be used:

- Logistic regression may be used when predicting whether bank customers are likely to default on their loans. This is a calculation a bank makes when deciding if it will or will not lend to a customer and assessing the maximum amount the bank will lend to those it has already deemed to be creditworthy. In order to make this calculation, the bank will look at several factors. Lend is the target in this logistic regression, and based on the likelihood of default that is calculated, a lender will choose whether to take the risk of lending to each customer.
    - These factors, also known as features or independent variables, might include credit score, income level, age, job status, marital status, gender, the neighborhood of current residence and educational history.
- Logistic regression is also often used for medical research and by insurance companies. In order to calculate cancer risks, researchers would look at certain patient habits and genetic predispositions as predictive factors. To assess whether or not a patient is at a high risk of developing cancer, factors such as age, race, weight, smoking status, drinking status, exercise habits, overall medical history, family history of cancer and place of residence and workplace, accounting for environmental factors, would be considered.
- Logistic regression is used in many other fields and is a common tool of data scient

Logistic regression assumptions

- Remove highly correlated inputs.
- Consider removing outliers in your training set because logistic regression will not give significant weight to them during its calculations.
- Does not favor sparse (consisting of a lot of zero values) data.
- Logistic regression is a classification model, unlike linear regression.

Maximum Likelihood Estimation

In statistics, maximum likelihood estimation (MLE) is **a method of estimating the parameters of an assumed probability distribution**, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.

As the name suggests in statistics it is a method for estimating the parameters of an assumed probability distribution. Where the likelihood function measures the goodness of fit of a statistical model on data for given values of parameters. The estimation of parameters is done by maximizing the likelihood function so that the data we are using under the model can be more probable for the model. The likelihood function for discrete random variables can be given by

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x),$$

Where x is the outcome of X random variables and likelihood is the function of θ. By the above function, we can say the likelihood is equal to the probability of occurrence of outcome x is observed when the parameter of the model is θ.

The likelihood function for continuous random variables can be given by

$$\mathcal{L}(\theta \mid x) = f_\theta(x),$$

Here the likelihood function can be put into hypothesis testing for finding the probability of various outcomes using the set of parameters defined in the null hypothesis.

The main goal of the maximum likelihood estimation is to make inferences about the data population which will take part in the generation of the sample and evaluating the joint density at the observed data set. As we have seen in the likelihood function above it can be maximized by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \widehat{L}_n(\theta\,; \mathbf{y})$$

Here the motive of the estimation is to select the best fit parameter for the model to make the data most probable. The specific value $\hat{\theta} = \hat{\theta}_n(\mathbf{y}) \in \Theta$ that maximizes the likelihood function Ln is called the maximum likelihood estimation.

Maximum Likelihood Estimation Vs. Least Square Method

The MLE is a "likelihood" maximization method, while OLS is a distance-minimizing approximation method. Maximizing the likelihood function determines the parameters that are most likely to produce the observed data. From a statistical point of view, MLE sets the mean and variance as parameters in determining the specific parametric values for a given model. This set of parameters can be used for predicting the data needed in a normal distribution.

Ordinary Least squares estimates are computed by fitting a regression line on given data points that has the minimum sum of the squared deviations (least square error). Both are used to estimate the parameters of a linear regression model. MLE assumes a joint probability mass function, while OLS doesn't require any stochastic assumptions for minimizing distance.

Maximum likelihood estimation, or MLE, is a method used in estimating the parameters of a statistical model, and for fitting a statistical model to data. If you want to find the height measurement of every basketball player in a specific location, you can use the maximum likelihood estimation. Normally, you would encounter problems such as cost and time constraints. If you could not afford to measure all of the basketball players' heights, the maximum likelihood estimation would be very handy. Using the maximum

likelihood estimation, you can estimate the mean and variance of the height of your subjects. The MLE would set the mean and variance as parameters in determining the specific parametric values in a given model.

To sum it up, the maximum likelihood estimation covers a set of parameters which can be used for predicting the data needed in a normal distribution. A given, fixed set of data and its probability model would likely produce the predicted data. The MLE would give us a unified approach when it comes to the estimation. But in some cases, we cannot use the maximum likelihood estimation because of recognized errors or the problem actually doesn't even exist in reality.

- ✓ "OLS" stands for "ordinary least squares" while "MLE" stands for "maximum likelihood estimation."
- ✓ The ordinary least squares, or OLS, can also be called the linear least squares. This is a method for approximately determining the unknown parameters located in a linear regression model.
- ✓ Maximum likelihood estimation, or MLE, is a method used in estimating the parameters of a statistical model and for fitting a statistical model to data.

Model Theory

Model Theory is the part of mathematics which shows how to apply logic to the study of structures in pure mathematics. On the one hand it is the ultimate abstraction; on the other, it has immediate applications to every-day mathematics.

The fundamental tenet of Model Theory is that mathematical truth, like all truth, is relative. A statement may be true or false, depending on how and where it is interpreted.
This isn't necessarily due to mathematics itself, but is a consequence of the language that we use to express mathematical ideas.

Model Theory is divided into two parts namely pure and applied. Pure model theory will learn the abstract properties of first order theories and there on derives structure theorems for their models. The applied model theory will study the concrete algebraic structures from model theoretic point of view and then uses the results from pure model theory functionalities and uniformities of definition. The applied model theory is connected strongly with other branches of mathematics.

The other areas of model theory are list out here.

1. Pure Model Theory
2. Model theory of fields with operators and connections with arithmetic Geometry.
3. Henselian Fields.
4. O-minimality and related Topics.
5. Model Theory of Groups.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Model Fit Statistics

- ✓ Model fit statistics describe and test the overall fit of the model.
- ✓ It measure the similarity between fitted model and actual outcome values that are generated.
- ✓ The problem of assessing model fit might be challenging if researchers tend to measure the fit that accounts for variability in model complexity, model misspecification and sample size.

Fit model describes the relationship between a response variable and one or more predictor variables.

There are many different models that you can fit including simple linear regression, multiple linear regression, analysis of variance (ANOVA), analysis of covariance (ANCOVA), and binary logistic regression.

Linear fit

A linear model describes the relationship between a continuous response variable and the explanatory variables using a linear function.

Logistic fit

A logistic model describes the relationship between a categorical response variable and the explanatory variables using a logistic function.

Mostly Fit Measures

- **Sum of Squared Errors (SSE):** Sum of squared differences between predicted and observed values. Measures deviation from actual values.
- **Log-likelihood (LL):** The Kullback-Leibler based measure of model fit to observed data. It will choose the model which can most likely create the in sample data.
- **Akaike Information Criterion (A/C):** A/C will allow the comparison between nested, overlapping or non nested models that have different numbers of parameters. It selects the model that makes out sample data most likely. It assumes that models are specified correctly.
- **Akaike Information Criterion with finite sample correction(A/CC):** The A/CC enables comparison between nested, overlapping or non nested models that have various number of parameters with less sample size correction. It assumes that models are correctly classified. Etc.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Construction

- ✓ Data modelling is the process of creating a data model for storing the data in the database
- ✓ A model is nothing but representation of data objects that is associations between different data objects and rules.
- ✓ Data Modelling helps in visually representing the data and the enforce business rules, regulatory compliances and the government policies on data.
- ✓ The logical designs are translated into physical models which contain storage devices, database and the files that build the data.. The business earlier used the relational database technology such as SQL to build the data models since it uniquely suits for linking the data set keys flexibility and data types to support the requirements of business processes.

An efficient model can be constructed by following steps..

- ✓ ***Do not impose traditional Modelling Techniques on Data***.
  Traditionally, fixed record data is stable and even predictable in its growth. With this the data modelling become easy. It sites contemplate the modelling of data. The effect of modelling must counter on building open and elastic data interfaces since users does not known when new data source or form of data can emerge.

- ✓ ***Design a system rather than Schema :*** in the era of traditional data, The relational database schema can cover the relationships and links between the data needed by business for its information support. This is not the case in big data that does not have database or that uses database such as NoSQL. The big data models must be created on systems rather than on databases.

- ✓ ***Use Data Modelling Tools:*** The IT decision makers must include the ability to create the data models for big data as the requirements while considering the big data tools and methodologies.
- ✓ ***Focus on the Core Data of Business:*** Enterprises get the data at large volumes. Most of the data is extraneous. The best method would be to identity the big data suitable for data tools and methodologies.

- ✓ ***Deliver the Quality Data:*** Earlier data models and relationships are effected for big data when organisations focus on development of sound definitions for data. The thorough meta data describes the source of data and its purpose. The knowledge about the data helps in planning it properly in data models to support the business.

- ✓ ***Search for Key in Roads into the data:*** A Commonly used vectors into big data today is geographical location. Based on the business the industries have other common keys into big data required by users. The data models can be created which support information access paths for the company by identifying the common entry points into the data.

Analytics Applications to Various Business Domains

The analytics applications in various business domains are illustrated as follows,

Digital Advertising: Data Algorithms can control the digital advertisements such as banners displayed on various websites to digital billboards in big cities.

Marketing: Analytics is used to observe the buying patterns of consumer behaviour, analyzing trends to identify the target audience through various advertising techniques which appeal to consumers, forecast supply needs etc.

Finance: Analytics is important to finance sector. The data scientists have high demand in investment banking portfolio management, financial planning, forecasting, budgeting etc.

CRM: Analytics enable to analyze the performance indicators that help in decision making and provide strategies boost the relationship with the consumers. The demographics and data about other socio economic factors, purchasing patterns, life cycle etc are important to CRM department.

Manufacturing: Analytics help in supply chain management, inventory management, measure the performance of targets, risk mitigation plans and even improve the efficiency based on product data.

Travel: Analytics help in optimization of travelers who buy the experience through social media and mobile mobile/weblog data analysis. The data analytics applications can deliver the personalized travel recommendations based on result from social media data.

***Customer interactions*:** Insurers can describe about their services through regular customer surveys after communicating with claim handlers. This is important to know about their goods.

***Manage risk*:** In insurance industry, the risk management is mainly focused. Data analytics offer insurance companies the data on claims, actuarial and risk data by covering the important decisions that the company must take. Evaluations are done by underwriter before anyone gets insured. Later on the appropriate insurance is set. Nowadays, analytical software are used to detect different fraudulent claims.

Delivery Logistics: Various logistic companies like UPS, DHL, FedEx etc, use data to improve their efficiency in operations. These companies from data analytics applications have found the suitable routes to ship, the best delivery time, suitable means of transport. Data generated by the companies through GPS provides them opportunities to take advantage of data analytics and data science.

Energy Management: Data analytics are applied to energy management and areas such as energy optimization, smart grid management, distribution of energy and building automation for utility companies are covered. The data analytics application focuses mainly on monitoring and controlling of dispatch crew, network devices and management of service outages.

HR Professionals: The HR Professionals use data to fetch information about educational background of skilled candidates, employee attrition rate, number of years of experiences service, age, gender etc. This data is useful to play pivotal role in candidate selection procedure.

Fraud and Risk Detection: Analytics helps to rescue from losses incurred by organizations since they could have extracted data from customers while applying loans. With this they can easily analyze and infer if there is any probability of customers defaulting.