Accuracy – Accuracy is a metric by which one can examine how good a machine learning model is. The confusion matrix to find accuracy of a model is the ratio of correctly

predicted classes to the total classes predicted.

True Positive + True negative / True Positive + True Negative + False Positive + False Negative

Adam Optimization – This algorithm is used in training deep learning models. It is an extension to stochastic gradient descent. In this optimization algorithm, running averages of both the gradients and the second moments of the gradients are used. It is used to compute adaptive learning rates for each parameter.

Features:

1. It is computationally efficient and has little memory requirements.
2. It is invariant to diagonal rescaling of the gradients
3. Adam works well in practice as compared to other stochastic optimization methods.

Apache spark – is an open source cluster computing framework. It can be deployed in variety of ways. So of its features are:

1. Speed – spark helps to run applications in Hadoop cluster, up to 100 times faster on memory and 10 times faster on disk.
2. Supports popular data science programming languages like R, Python and Scala.
3. Spark also has a library called MLIIB which has some basic machine learning models.

Auto Regression – It is a time series model that uses previous time steps as an input to equation to predict the value at the next time step.

* X(t+1) = b0 + b1\*X(t-1) + b2\*X(t-2)

Since the regression model uses data from the same input variable at previous time steps, it is referred to as an autoregression.

Back Propagation – In neural networks, if the estimated output is far away from the actual output (high error), we update the biases and weights based on the error. This weight and bias updating process is known as Back Propagation. Back-propagation (BP) algorithms work by determining the loss (or error) at the output and then propagating it back into the network

Bagging – Bagging or bootstrap averaging is a technique where multiple models are created on the subset of data, and the final predictions are determined by combining the predictions of all the models. Some of the algorithms that use bagging technique are : Bagging meta-estimator, Random Forest.

Bar Chart – Bar charts are a type of graph that are used to display and compare the numbers, frequency or other measures (e.g. mean) for different discrete categories of data. They are used for categorical variables.

Bayes Theorem – Bayes’ theorem is used to calculate the conditional probability. Conditional probability is the probability of an event ‘B’ occurring given the related event ‘A’ has already occurred.

Formula : P(A/B) = P(B/A)P(A) / P(B)

Bayesian Statistics – It is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data. It differs from classical frequentist approach and is based on the use of Bayesian probabilities to summarize evidence.

1. Bias Variance Trade Off - The error emerging from any model can be broken down into components mathematically. **Bias error** is useful to quantify how much on an average are the predicted values different from the actual value
2. **Variance** on the other side quantifies how are the prediction made on same observation different from each other.

A high bias error means we have a under-performing model which keeps on missing important trends. A high variance model will over-fit on your training population and perform badly on any observation beyond training. In order to have a perfect fit in the model, the bias and variance should be balanced which is bias variance trade off.

Big Data – Big data is a term that describes the large volume of data – both structured and unstructured. organizations use this large amount of data to generate insights. Companies use various tools, techniques and resources to make sense of this data to derive effective business strategies.

Binary Variable – Binary variables are those variables which can have only two unique values.

Binomial Distribution – It is a method of calculating probabilities for experiments having fixed number of trials. Distribution is applied only on discrete random variable.

Binomial distribution has following properties:

1. The experiment should have finite number of trials
2. There should be two outcomes in a trial: success and failure
3. Trials are independent
4. Probability of success (p) remains constant

Formula – P(X = r) = nCr(P^r)\*(1-P)\*(n-r)

Boosting – Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Some of the boosting algorithms are: AdaBoost, GBM, XGBM, LightGBM, CatBoost

Bootstrapping – It is the process of dividing the dataset into multiple subsets, with replacement. Each subset is of the same size of the dataset. These samples are called bootstrap samples.

Box Plot – It displays the full range of variation (from min to max), the likely range of variation (the Interquartile range*)*, and a typical value (the median).

Business Analytics – Business analytics is mainly used to show the practical methodology followed by an organization for exploring data to gain insights. The methodology focusses on statistical analysis of the data.

Business Intelligence – Business intelligence are a set of strategies, applications, data, technologies used by an organization for data collection, analysis and generating insights to derive strategic business opportunities.

Categorical Value –  Categorical variables (or nominal variables) are those variables which have discrete qualitative values. For example, names of cities are categorical like Delhi, Mumbai, Kolkata.

Classification – It is supervised learning method where the output variable is a category, such as “Male” or “Female” or “Yes” and “No”.

 Classification Threshold – Classification threshold is the value which is used to classify a new observation as 1 or 0. When we get an output as probabilities and have to classify them into classes, we decide some threshold value and if the probability is above that threshold value we classify it as 1, and 0 otherwise.

Clustering – Clustering is an unsupervised learning method used to discover the inherent groupings in the data.

Computer vision – Computer Vision is a field that deals with enabling computers to visualize, process and identify images/videos in the same way that a human vision does.

Concordant – discordant ratio – Concordant and discordant pairs are used to describe the relationship between pairs of observations. To calculate the concordant and discordant pairs, the data are treated as ordinal. The number of concordant and discordant pairs are used in calculations for Kendall’s tau, which measures the association between two ordinal variables.

Confidence Interval – A confidence interval is used to estimate what percent of a population fits a category based on the results from a sample population.

Confusion Matrix – A confusion matrix is a table that is often used to describe the performance of a classification model. It is a N \* N matrix, where N is the number of classes. We form confusion matrix between prediction of model classes Vs actual classes. The 2nd quadrant is called type II error or False Negatives, whereas 3rd quadrant is called type I error or False positives

Continuous Variable – Continuous variables are those variables which can have infinite number of values but only in a specific range. For example, height is a continuous variable.

Convergence – Convergence refers to moving towards union or uniformity. An iterative algorithm is said to converge when as the iterations proceed the output gets closer and closer to a specific value.

Convex Function – A real value function is called convex if the line segment between any two points on the graph of the function lies above or on the graph.

Correlation - Correlation is the ratio of covariance of two variables to a product of variance (of the variables). It takes a value between +1 and -1. An extreme value on both the side means they are strongly correlated with each other. A value of zero indicates a NIL correlation but not a non-dependence.

The most widely used correlation coefficient is Pearson Coefficient.

Cosine Similarity – Cosine Similarity is the cosine of the angle between 2 non-zero vectors. Two parallel vectors have a cosine similarity of 1 and two vectors at 90° have a cosine similarity of 0.

Cost Function – Cost function is used to define and measure the error of the model.

[EX : So let’s say, you increase the size of a particular shop, where you predicted that the sales would be higher. But despite increasing the size, the sales in that shop did not increase that much. So the cost applied in increasing the size of the shop, gave you negative results. So, we need to minimize these costs. Therefore we make use of cost function to minimize the loss.]

Covariance – Covariance is a measure of the joint variability of two random variables. It’s similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

Cross Entropy – The cross entropy between two probability distributions and over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set, if a coding scheme is used that is optimized for an “unnatural” probability distribution , rather than the “true”. Cross entropy can be used to define the loss function in machine learning and optimization.

Cross Validation – It is a technique which involves reserving a particular sample of a dataset which is not used to train the model. Later, the model is tested on this sample to evaluate the performance. There are various methods of performing cross validation such as:

1. Leave one out cross validation (LOOCV)
2. k-fold cross validation
3. Stratified k-fold cross validation
4. Adversarial validation

Data Mining – Data mining is a study of extracting useful information from structured/unstructured data taken from various sources. This is done usually for purposes like Market Analysis, determining customer purchase pattern, financial planning, fraud detection, etc.,

Data Science – Data science is a combination of data analysis, algorithmic development and technology in order to solve analytical problems. The main goal is a use of data to generate business value.

Data Transformation – Data transformation is the process to convert data from one form to the other.

Database – Database (abbreviated as DB) is an structured collection of data. The collected information is organized in a way such that it is easily accessible by the computer. Databases are built and managed by using database programming languages. The most common database language is SQL.

Dataframe – Dataframe is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects.

Dataset – A dataset is a collection of data. A dataset is organized into some type of data structure. everal characteristics define a dataset’s structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis.

Dashboard – Dashboard is an information management tool which is used to visually track, analyze and display key performance indicators, metrics and key data points. Popular tools for building dashboards include Excel and Tableau.

DBscan – DBSCAN is the acronym for **Density-Based Spatial Clustering of Applications with Noise**. It is a clustering algorithm that isolates different density regions by forming clusters. For a given set of points, it groups the points which are closely packed. The algorithm has two important features: Distance, The minimum number of points required to form a dense region.

Decision Boundary – In a statistical-classification problem with two or more classes, a decision boundary or decision surface is a hypersurface that partitions the underlying vector space into two or more sets, one for each class. How well the classifier works depends upon how closely the input patterns to be classified resemble the decision boundary.

Decision Tree – Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input & output variables. In this technique, we split the population (or sample) into two or more homogeneous sets based on most significant splitter / differentiator in input variables.

Deep Learning – It is associated with a machine learning algorithm (Artificial Neural Network, ANN) which uses the concept of human brain to facilitate the modeling of arbitrary functions. ANN requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously.

Descriptive Statistics – It is comprised of those values which explains the spread and central tendency of data.

Decile – Decile divides a series into 10 equal parts. For any series, there are 10 decile denoted by D1, D2, D3 … D10. These are known as First Decile , Second Decile and so on

Degree Of Freedom – It is the number of variables that have the choice of having more than one arbitrary value.

Dimensionality Reduction – Dimensionality Reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. Dimension Reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely.

Dplyr – Dplyr is a popular data manipulation package in R. It makes data manipulation, cleaning, summarizing very user friendly. Dplyr can work not only with the local datasets, but also with remote database tables, using exactly the same R code

Dummy Variable – Dummy Variable is another name for Boolean variable. An example of dummy variable is that it takes value 0 or 1. 0 means value is true (i.e. age < 25) and 1 means value is false (i.e. age >= 25)

Early stopping – Early stopping is a technique for avoiding overfitting when training a machine learning model with iterative method. We set the early stopping in such a way that when the performance has stopped improving on the held-out validation set, the model training stops.

EDA – EDA or exploratory data analysis is a phase used for data science pipeline in which the focus is to understand insights of the data through visualization or by statistical analysis.

ETL – ETL is the acronym for Extract, Transform and Load. It has the following properties:

1. It extracts data from the source systems
2. It enforces data quality and consistency standards
3. Delivers data in a presentation-ready format

This data can be used by application developers to build applications and end users for making decisions

Evaluation Metrics – The purpose of evaluation metric is to measure the quality of the statistical / machine learning model. AUC, ROC score, F-Score, Log-Loss are some of the evaluation metrics.

Factor Analysis – Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. Factor analysis aims to find independent latent variables. Factor analysis also assumes several assumptions:

* There is linear relationship
* There is no multicollinearity
* It includes relevant variables into analysis
* There is true correlation between variables and factors

There are different types of methods used to extract the factor from the data set:

1. Principal Component Analysis
2. Common factor analysis
3. Image factoring
4. Maximum likelihood method

False Negative – Points which are actually true but are incorrectly predicted as false.

False Positive – Points which are actually false but are incorrectly predicted as true.

Feature hashing – It is a method to transform features to vector. Without looking up the indices in an associative array, it applies a hash function to the features and uses their hash values as indices directly.

Feature reduction – Feature reduction is the process of reducing the number of features to work on a computation intensive task without losing a lot of information.

Feature selection – Feature Selection is a process of choosing those features which are required to explain the predictive power of a statistical model and dropping out irrelevant features.

Few-shot learning – It refers to the training of machine learning algorithms using a very small set of training data instead of a very large set. This is most suitable in the field of computer vision, where it is desirable to have an object categorization model work well without thousands of training examples.

Flume – Flume is a service designed for streaming logs into the Hadoop environment. It can collect and aggregate huge amounts of log data from a variety of sources. In order to collect high volume of data, multiple flume agents can be configured.

Frequentist Statistics – Ittests whether an event (hypothesis) occurs or not. It calculates the probability of an event in the long run of the experiment (i.e the experiment is repeated under the same conditions to obtain the outcome).

F-score – It is an evaluation metric combines both precision and recall as a measure of effectiveness of classification. It is calculated in terms of ratio of weighted importance on either recall or precision as determined by β coefficient.

F measure = 2 x (Recall × Precision) / ( β² × Recall + Precision )