Q1. What is Subquery?

Ans: A subquery is a query within another query. The outer query is called as main query, and inner query is called subquery. Subquery is always executed first, and the result of subquery is passed on to the main query.

There are two types of Subqueries.

1. Correlated Subquery

2.Non-Correlated Subquery

Q2. What is the difference between DELETE and TRUNCATE commands?

Ans: DELETE command is used to remove rows from the table, and WHERE clause can be used for conditional set of parameters. Commit and Rollback can be performed after delete statement.

TRUNCATE removes all rows from the table. Truncate operation cannot be rolled back.

Q3. What is the difference between 'BETWEEN' and 'IN' condition operators?

Ans: BETWEEN operator is used to display rows based on a range of values in a row whereas the IN-condition operator is used to check for values contained in a specific set of values.

Q4. What is the difference between Cluster and Non- cluster index?

Ans: Clustered index is used for easy retrieval of data from the database by altering the way that the records are stored. Database sorts out rows by the column which is set to be clustered index.

A non-clustered index does not alter the way it was stored but creates a complete separate object within the table. It points back to the original table rows after searching.

Q1. What is the difference between Stochastic gradient descent (SGD) and Gradient Descent (GD)?

Ans: Gradient Descent and Stochastic gradient are the algorithms that find the set of parameters that will minimize a loss function.

In Gradient Descent (GD), we perform the forward pass using ALL the train data before starting the backpropagation pass to adjust the weights.

In Stochastic Gradient Descent (SGD), we perform the forward pass using a SUBSET of the train set followed by backpropagation to adjust the weights.

Q2. What is the difference between normalization and standardization? When to use normalization & standardization?

Ans: Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between O and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Q3. What is pooling & its types?

Ans: Pooling layers are used to reduce the dimensions of the feature maps. It reduces the number of parameters to learn and the amount of computation performed in the network.

The pooling layer summarizes the features present in a region of the feature map generated by a convolution layer.

Types of pooling:

1.Max Pooling 2. Average pooling 3. Global Pooling

Q4. Explain concepts of RNN?

Ans: RNN is the recurrent neural network which is used to analyze and recognize the pattern in the sequence of data. Due to their internal memory, RNN can certainly remember the things about the inputs they receive.

Most common issues faced with RNN: Exploding gradients and Vanishing gradients

Q1. What is Linear Regression?

Ans: Linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

There major uses for Regression Analysis are:

1.Determining the Strength of Predictors

2.Forecasting an Effect

3.Trend Forecasting

Q2. How we can Calculate Error in Linear Regression?

Ans: Linear regression most often uses mean-square-error (MES) to calculate the error of the model. MSE is Calculated by:

\* Measuring the distance of the observed y-values from the predicted y-values at each value of X. \* Squaring each of these distances. \* Calculating the mean of each of the squared distance.

Q3. Difference between loss and cost function?

Ans: Loss Function: A loss function is a function measure the error between a single prediction and the corresponding actual value.

Cost Function: A cost function is a function which measure the error between predictions and their actual values across the whole dataset.

Both the cost and the loss function calculate prediction error, but the key difference is the level at which they calculate this error. A lost function calculates the error per observation, while the lost function calculates the error for all observations by aggregating the loss values.

Q4. Difference between linear & logistic regression?

Ans: Linear and Logistic regression are the most basic form of regression which are commonly used. The main difference between these two algorithms is that Logistic regression is used when the dependent variable is binary. In contrast, Linear regression is used when the dependent variable is continuous, and the nature of the regression line is linear.

Linear regression requires to establish the linear relationship among dependent and independent variables, whereas it is not necessary for logistic regression.

Q1. What is p-value?

Ans: p-value helps you determine the strengths of your results when you perform a hypothesis test. It is a number between 0 and 1. The claim which is on trial is called the Null Hypothesis. Lower p-values, i.e., < = 0.05, means we can reject the Null Hypothesis. A high p-value, i.e., >= 0.05, means we can accept the Null Hypothesis. An exact p-value 0.05 indicates that the Hypothesis can go either way.

Q2. Which are the important steps of Data Cleaning?

Ans: Different types of data require different types of cleaning; the most important steps of Data Cleaning are:

1.Data Quality

2.Removing Duplicate Data (also irrelevant data)

3.Structural errors

4.Outliers

5.Treatment for Missing Data

Q3. Explain Normal Distribution

Ans: Normal Distribution is also called the Gaussian Distribution. It is a type of probability distribution such that most of the values lie near the mean. It has the following t

characteristics:

1. The mean, median, and mode of the distribution coincide
2. The distribution has a bell-shaped curve
3. The total area under the curve is 1 Exactly half of the values are to the right of the center,

and the other half to the left of the center

Q4. What is correlation and covariance in statistics?

Ans: Correlation is defined as the measure of the relationship between two variables. If two variables are directly proportional to each other, then its positive correlation. If the variables are indirectly proportional to each other, it is known as a negative correlation.

Covariance is the measure of how much two random variables vary together.

Q1. What is Cross-Validation?

Ans: It is a model validation technique to assess how the outcomes of a statistical analysis will infer to an independent data set. It is majorly used where prediction is the goal and one needs to estimate the performance accuracy of a predictive model in practice.

The goal here is to define a data-set for testing a model in its training phase and limit overfitting and underfitting issues. The validation and the training set is to be drawn from the same distribution to avoid making things worse.

Q2. What is the importance of A/B testing?

Ans: The goal of A/B testing is to pick the best variant among two hypotheses, the use cases of this kind of testing could be a web page or application responsiveness, landing page redesign, banner testing, marketing campaign performance etc. The first step is to confirm a conversion goal, and then statistical analysis is used to understand which alternative performs better for the given conversion goal.

Q3. Explain Eigenvectors and Eigenvalues

Ans: Eigenvectors depict the direction in which a linear transformation moves and acts by compressing, flipping, or stretching. They are used to \_ understand linear transformations and are generally calculated for a correlation or covariance matrix. the eigenvalue is the strength of the transformation in the direction of the eigenvector. An eigenvector’s direction remains unchanged when a linear transformation is applied to it.

Q4. How should you maintain a deployed model?

Ans: A deployed model needs to be retrained after a while so as to improve the performance of the model. Since deployment, a track should be kept of the predictions made by the model and the truth values. Later this can be used to retrain the model with the new data. Also, root cause analysis for wrong predictions should be done.

Q1. What is dimensionality reduction?

Ans: Dimensionality reduction is the process of converting a dataset with a high number of dimensions (fields) to a dataset with a lower number of dimensions. This is done by dropping some fields or columns from the dataset. However, this is not done haphazardly. In this process, the dimensions or fields are dropped only after making sure that the remaining information will still be enough to succinctly describe similar information.

Q2. Explain the steps in making a decision tree.

1.Take the entire data set as input

2.Calculate entropy of the target variable, as well as the predictor attributes

3.Calculate your information gain of all attributes (we gain information on sorting different objects from each other)

4.Choose the attribute with the highest information gain as the root node

5.Repeat the same procedure on every branch until the decision node of each branch is finalized

Q3. What is Linear Regression? What are some of the major drawbacks of the linear model?

Ans: Linear regression is a technique in which the score of a variable Y is predicted using the score of a predictor variable X. Y is called the criterion variable. Some of the drawbacks of Linear Regression are as follows:

The assumption of linearity of errors is a major drawback. It cannot be used for binary outcomes. We have Logistic Regression for that. Overfitting problems are there that can’t be solved.

Regression is a type of Machine learning which helps in finding the relationship between

independent and dependent variable. In simple words, Regression can be defined as a

Machine learning problem where we have to predict discrete values like price, Rating, Fees, etc.

Evaluation Metrics for Regression

1.Mean Absolute Error (MAE) 2. Mean Squared Error (MSE) 3. Root Mean Squared Error (RMSE) 4. R Squared (R2) 5. Adjusted R Squared

Mean Absolute Error (MAE),

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

Mean Squared Error (MSE)

MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value. In MAE, we are finding the absolute difference and here we are finding the squared difference.

Root Mean Squared Error (RMSE)

As RMSE is clear by the name itself, that it isa simple square root of mean squared error.

R Squared (R2)

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

Ensemble learning helps to enhance the performance of machine learning models. The concept behind it is simple. Multiple machine learning models are combined to obtain a more accurate model.

Bagging, boosting and stacking are the three most popular ensemble learning techniques. Each of these techniques offers a unique approach to improving predictive accuracy. Each technique is used fora different purpose, with the use of each depending on varying factors. Although each technique is different, many of us find it hard to distinguish between them. Knowing when or why we should use each technique is difficult.

Ensemble learning is a learning method that consists of combining multiple machine learning models.

A problem in machine learning is that individual models tend to perform poorly. In other words, they tend to have low prediction accuracy. To mitigate this problem, we combine multiple models to get one with a better performance.

the individual models that we combine are known as weak learners. We call them weak learners because they either have a high bias or high variance.

Both high bias and high variance models thus cannot generalize properly. Thus, weak learners will either make incorrect generalizations or fail to generalize altogether. Because of this, the predictions of weak learners cannot be relied on by themselves.

As we know from the bias-variance trade-off, an underfit model has high bias and low variance, whereas an overfit model has high variance and low bias. In either case, there is no balance between bias and variance. For there to be a balance, both the bias and variance need to be low. Ensemble learning tries to balance this bias-variance trade-off by reducing either the bias or the variance.

Ensemble learning will aim to reduce the bias if we have a weak model with high bias and low variance. Ensemble learning will aim to reduce the variance if we have a weak model with high variance and low bias. This way, the resulting model will be much more balanced, with low bias and variance. Thus, the resulting model will be known as a strong learner. This model will be more generalized than the weak learners. It will thus be able to make accurate predictions.

Ensemble learning improves a model's performance in mainly three

1. By reducing the variance of weak learners 2. By reducing the bias of weak learners,

3. By improving the overall accuracy of strong learners.

What is the difference between normalization and standardization? When to use normalization & standardization?

What is the meaning of Overfitting in Machine learning?

Why does overfitting occur?

What is the method to avoid overfitting?

How does Machine Learning differ from Deep Learning?

How is KNN different from k-means?

What are the five popular algorithms we use in Machine Learning?

What is model selection in Machine Learning?

What are the common ways to handle missing data in ds

Describe Precision and Recall?

Explain True Positive, True Negative, False Positive, and False Negative in Confusion

What, according to you, is more important between model accuracy and model performance?

What do you understand by the F1 score?

How is a decision tree pruned?

What do you understand by Underfitting?

When does regularization become necessary in Machine learning?

What is Regularization? What kind of problems does regularization solve?

Do you think that treating a categorical variable as a continuous variable would result in a better predictive model?

How is machine learning used in day-to-day life?

Describe dimension reduction in machine learning.

What is the trade-off between bias and variance?

What do you understand about the Decision Tree in Machine Learning?

What do you mean by Genetic Programming?

Why do we need to convert categorical variables into factors? Which functions are used to perform the conversion?

How Can You Choose a Classifier Based on a Training set data set?

Explain the Confusion Matrix with Respect to Machine Lea algo

What Is a False Positive and False Negative and How are significant?

How is Flipkart Able to Recommend Other Things to Buy? How Does the Recommendation Engine Work?

Briefly Explain Logistic Regression.

What Are Some Methods of Reducing Dimensionality?

Explain Correlation and Covariance?

What are Loss Function and Cost Functions? Explain

How do you make sure which Machine Learning algorithms to use?

How do you check the Normality of a dataset?

What is P-value?

How can we relate standard deviation and variance?

Is a high variance in data good or bad?

What is a Box-Cox transformation?

What is the difference between regularization and normalization?

Explain the difference between L1 and L2

How would you handle an imbalanced dataset?

What is the difference between Gini Impurity and Entropy in a Decision Tree?