Company Name - Cognizant

Role - Data Scientist

Q. Describe how Gradient Boosting works?

A. Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. If a small change in the prediction for a case causes no change in error, then next target outcome of the case is zero. Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Q. Describe the decision tree model?

A. Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The leaves are the decisions or the final outcomes. A decision tree is a machine learning algorithm that partitions the data into subsets.

Q. What is a neural network?

A. Neural networks are a set of algorithms, modelled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labelling or clustering raw input. They, also known as Artificial Neural Networks, are the subset of Deep Learning.

Q. Explain the Bias-Variance Tradeoff?

A. The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

Q. Fix multi-collinearity in a regression model?

A. Follow these methods:

1. Remove some of the highly correlated independent variables.
2. Linearly combine the independent variables, such as adding them together.
3. Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

Q. Differentiate between gradient boosting and random forest?

A. The two main differences are: How trees are built: random forests builds each tree independently while gradient boosting builds one tree at a time. Combining results: random forests combine results at the end of the process (by averaging or "majority rules") while gradient boosting combines results along the way.

Q. Can you write a user-defined function in SQL? how?

A. SQL Server allows us to create our functions called as user defined functions in SQL Server. For example, if we want to perform some complex calculations, then we can place them in a separate function, and store it in the database. Whenever we need the calculation, we can call it.

Q. Hash tables? Where do you use it?

A. A hash table is a data structure that is used to store keys/value pairs. It uses a hash function to compute an index into an array in which an element will be inserted or searched. They are widely used in many kinds of computer software, particularly for associative arrays, database indexing, caches and sets. The idea of a hash table is to provide a direct access to its items. So that is why the it calculates the "hash code" of the key and uses it to store the item, instead of the key itself.

Q. in a scenario, what would you prioritize: bias or variance?

A. Bias is an error between the actual values and the model's predicted values. Variance is also an error but from the model's sensitivity to the training data. A prioritization of Bias over Variance will lead to a model that overfits the data. Prioritizing Variance will have a model underfit the data.

Q. Probabilistic and statistical modelling?

A. Probabilistic modelling is a statistical technique used to take into account the impact of random events or actions in predicting the potential occurrence of future outcomes. Statistical modelling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.

Q. What is Scaling? It's advantage?

A. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

Q. Mix-Max Normalization vs Standardization?

A. Min-Max Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Q. Describe Feature engineering?

A. Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

Q. Left and right Skewness?

A. Left Skewed Distribution: In a left skewed distribution, the mean is less than the median. Right Skewed Distribution: In a right skewed distribution, the mean is greater than the median.

Q. What is Unsupervised algorithms?

A. Unsupervised learning aims to discover the dataset's underlying pattern, assemble that data according to similarities, and express that dataset in a precise format. Example: K-means clustering, anomaly detection, PCA etc

---