Company Name - Mindtree
Role - Data Scientist

Q. What is Central tendency?

A. Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution."[2] It aims to provide an accurate description of the entire data. It is the single value that is most typical/representative of the collected data.

Q. What is Central limit theorem?

A. The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

Q. What is Chi-Square test?

A. A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

Q. What is A/B testing?

A. A/B testing is an optimisation technique often used to understand how an altered variable affects audience or user engagement. It's a common method used in marketing, web design, product development, and user experience design to improve campaigns and goal conversion rates.

Q. What is Outlier treatment method?

A. Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process. If the outlier does not change the results but does affect assumptions, you may drop the outlier. Or just trim the data set, but replace outliers with the nearest "good" data, as opposed to truncating them completely.

Q. Describe ANOVA test?

A. Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
ANOVA checks the impact of one or more factors by comparing the means of different samples.

Q. What is Cross validation?

A. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

Q. How will you work in a machine learning project if there is a huge imbalance in the data?

A. Follow these techniques:

1. Use the right evaluation metrics.
2. Use K-fold Cross-Validation in the right way.
3. Ensemble different resampled datasets.
4. Resample with different ratios.
5. Cluster the abundant class.
6. Design your own models.

Q. Formula of sigmoid function?

A. It is a mathematical function having a characteristic that can take any real value and map it to between 0 to 1 shaped like the letter "S".

Y = 1 / 1+(e*-z)

Q. Which metric is used to split a node in Decision Tree?

A. The Gini Index and the Entropy and Information gain metrics are the metrics to use in the algorithm to create a decision tree

Q. What is ensemble learning?

A. Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

Q. Shallow copy and deep copy?

A. In Shallow copy, a copy of the original object is stored and only the reference address is finally copied. In Deep copy, the copy of the original object and the repetitive copies both are stored. Shallow copy is faster than Deep copy. The changes made in the copied object also reflect the original object in shallow copy. There is no reflection on the original object when the changes are made in the copied object in deep copy.

Q. What is Dimension reduction?

A. Dimensionality Reduction is used to reduce the feature space with consideration by a set of principal features.

Q. What is precision/recall ratio?

A. When it comes to precision we're talking about the true positives over the true positives plus the false positives. As opposed to recall which is the number of true positives over the true positives and the false negatives.

Q. Recommender Systems?

A. The recommender system mainly deals with the likes and dislikes of the users. Its major objective is to recommend an item to a user which has a high chance of liking or is in need of a particular user based on his previous purchases. It is like having a personalized team who can understand our likes and dislikes and help us in making the decisions regarding a particular item without being biased by any means by making use of a large amount of data in the repositories which are generated day by day.
