**Company Name - Larsen and Toubro**
**Role - Data Scientist**

### Q. What args will return?

A. The special syntax *args in function is used to pass a variable number of arguments to a function. It is used to pass a non-key worded, variable-length argument list. The syntax is to use the symbol * to take in a variable number of arguments; by convention, it is often used with the word args.

### Q. Difference between having and where clause in SQL?

A. WHERE Clause is used to filter the records from the table based on the specified condition. HAVING Clause is used to filter record from the groups based on the specified condition.

### Q. Assumptions in Multiple linear regression?

A. The regression has five key assumptions:

- Linear relationship.
- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity.

### Q. what is Gini index?

A. Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.

### Q. what is Entropy?

A. Entropy is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty. Entropy can be defined as a measure of the purity of the sub split. Entropy always lies between 0 to 1.

### Q. Describe Random forest algorithm?

A. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on model ensemble learning technique.

### Q. Tell us something about XGBoost Algorithm?

A. XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

**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 VIF?**

A. Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

**Q. Difference Between Bagging and Boosting?**

A. Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

**Q. P value and it's significance?**

A. The p-value is the probability that the null hypothesis is true. (1 – the p-value) is the probability that the alternative hypothesis is true. A low p-value shows that the results are replicable. A low p-value shows that the effect is large or that the result is of major theoretical, clinical or practical importance.

**Q. what is Type 1 and Type II error?**

A. A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

**Q. What are the Ways to avoid overfitting?**

A. Some steps that we can take to avoid it:

1. Data augmentation
2. L1/L2 Regularization
3. Remove layers / number of units per layer
4. Cross-validation

Q. What Is Image classification algorithms

A. Image Classification algorithms are the algorithms which are used to classify labels for images using their characteristics. Example: Convolutional Neural Networks.

**Q. Handle data with lot of noise?**

A. Noisy data can be handled by following the given procedures:
1) Binning:
• Binning methods smooth a sorted data value by consulting the values around it.
• The sorted values are distributed into a number of “buckets,” or bins.
• Because binning methods consult the values around it, they perform local smoothing.
• Similarly, smoothing by bin medianscan be employed, in which each bin value is replaced by the bin median.
• In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries.
• Each bin value is then replaced by the closest boundary value.
• In general, the larger the width, the greater the effect of the smoothing.
• Alternatively, bins may be equal-width, where the interval range of values in each bin is constant.
• Binning is also used as a discretization technique.
2) Regression:
• Here data can be smoothed by fitting the data to a function.
• Linear regression involves finding the “best” line to fit two attributes, so that one attribute can be used to predict the other.
• Multiple linear regressions an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.
3) Clustering:
• Outliers may be detected by clustering, where similar values are organized into groups, or “clusters.”
• Similarly, values that fall outside of the set of clusters may also be considered outliers.



**Q. Min sample leaf and max depth in random forest?**

A. Min sample leaf specifies the minimum number of samples that should be present in the leaf node after splitting a node. The max_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node. Using the max_depth parameter, I can limit up to what depth I want every tree in my random forest to grow.

**Q. Dict and list comprehension?**

A. The only difference between list and dictionary comprehension is that the dictionary has the keys and values in it. So, the keys and values will come as the expression or value.

**Q. How do you handle categorical data?**

A. One-Hot Encoding is the most common, correct way to deal with non-ordinal categorical data. It consists of creating an additional feature for each group of the categorical feature and mark each observation belonging (Value=1) or not (Value=0) to that group.

**Q. What Is Interpolation and Extrapolation?**

A. Interpolation is the process of calculating the unknown value from known given values whereas extrapolation is the process of calculating unknown values beyond the given data points.

**Q. Tell us something about SQL joins and Groups?**

A. The SQL Joins clause is used to combine records from two or more tables in a database. The GROUP BY statement groups rows that have the same values into summary rows, like "find the number of customers in each country".

**Q. How do you handle null values and which Imputation method is more favourable?**

A. Ways to handle missing values in the dataset:

1. Deleting Rows with missing values.
2. Impute missing values for continuous variable.
3. Impute missing values for categorical variable.
4. Other Imputation Methods.
5. Using Algorithms that support missing values.
6. Prediction of missing values.

Multiple imputation is more advantageous than the single imputation because it uses several complete data sets and provides both the within-imputation and between-imputation variability.

---

**NOTE:**
If you want to learn Data Science from scratch and become interview ready, check out our popular Data science course.

https://ai.cloudyml.com/Learn-Data-Science-from-Scratch