Company Name - TCS
Role - Data Scientist

Q. What is an imbalanced data set?

A. Imbalanced data sets are a special case for classification problem where the class distribution is not uniform among the classes.

Q. What are the approaches for treating the missing values?

A. They are:
List wise or case deletion
Pairwise deletion
Mean substitution
Regression imputation
Maximum likelihood.

Q. Evaluation metrics for Classification?

A. The key classification metrics are Accuracy, Recall, Precision, and F1- Score.

Q. Bagging vs Boosting with examples

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.
If we have 5 bagged decision trees that made the following class predictions for an input sample: blue, blue, red, blue and red, we would take the most frequent class and predict blue. This is an example of bagging algorithm.
While AdaBoost (Adaptive Boosting), Gradient Tree Boosting, XGBoost are the examples of boosting algorithms.

Q. Handling of imbalanced datasets?

A. Techniques to Handle Imbalanced Data
Use the right evaluation metrics
Use K-fold Cross-Validation in the right way
Ensemble different resampled datasets
Resample with different ratios
Cluster the abundant class
Design your own models

Q. Order of execution of SQL?

A. The order of execution is
1. FROM 2. JOIN 3. WHERE 4. GROUP BY 5. HAVING 6. SELECT 7. ORDER BY

Q. How can you assess a good logistic model?

A. Measuring the performance of Logistic Regression:
1. One can evaluate it by looking at the confusion matrix and count the misclassifications (when using some probability value as the cutoff)
2. One can evaluate it by looking at statistical tests such as the Deviance or individual Z-scores.

Q. What is bias?

A. The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.

Q. Ensemble Learning?

A. Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.)

Q. Normal Distribution? Skewed distribution? Solution?

A. The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one. If one tail is longer than another, the distribution is skewed. These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry. Symmetry means that one half of the distribution is a mirror image of the other half. Dealing with skew data: log transformation: transform skewed distribution to a normal distribution:
1. Remove outliers.
2. Normalize (min-max)
3. Cube root: when values are too large.
4. Square root: applied only to positive values.
5. Reciprocal.
6. Square: apply on left skew.

Q. capture the correlation between continuous and categorical variable? How?

A. There are three big-picture methods to understand if a continuous and categorical are significantly correlated - point biserial correlation, logistic regression, and Kruskal Wallis H Test. The point biserial correlation coefficient is a special case of Pearson's correlation coefficient.

Q. variance? Is it good or bad in data?

A. Variance refers to the changes in the model when using different portions of the training data set. Simply stated, variance is the variability in the model prediction—how much the ML function can adjust depending on the given data set.

Q. error and a residual error?

A. The error (or disturbance) of an observed value is the deviation of the observed value from the (unobservable) true value of a quantity of interest (for example, a population mean), and the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest.

Q. select an appropriate value of k in k-means?

A. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster.

Q. dictionaries or lists faster for lookups?

A. Lookups are faster in dictionaries because Python implements them using hash tables. If we explain the difference by Big O concepts, dictionaries have constant time complexity, O(1) while lists have linear time complexity, O(n).

Q. How Regularly Must an Algorithm be Updated?

A. It can vary time to time depending upon number of updates happened in the algorithm as per the requirement.

Q. Why Is Resampling Done?

A. Resampling methods are used to ensure that the model is good enough and can handle variations in data. The model does that by training it on the variety of patterns found in the dataset.

Q. Write the formula to calculate R-square?

A. R^2 = 1 - (RSS/TSS) where RSS = sum of squares of residual and TSS = Total sum of squares

Q. Goal of A/B Testing?

A. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

Q. Do Gradient Descent Methods at All-Time Converge to a Similar Point?

A. No, they always don't. That's because in some cases it reaches a local minima or a local optima point.

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