Company Name - Infosys
Role - Data Scientist

Q. curse of dimensionality? How would you handle it?

A. To overcome the issue of the curse of dimensionality, Dimensionality Reduction is used to reduce the feature space with consideration by a set of principal features.

Q. How to find the multi collinearity in the data set?

A. A simple method to detect multicollinearity in a model is by using something called the variance inflation factor or the VIF for each predicting variable.

Q. Explain the difference ways to treat multi collinearity!

A. 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. How you decide which feature to keep and which feature to eliminate after performing multi collinearity test?

A. It is advisable to get rid of variables iteratively. We would begin with a variable with the highest VIF score since other variables are likely to capture its trend. As a result of removing this variable, other variables' VIF values are likely to reduce.

Q. Explain logistic regression?

A. Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

Q. P value and its significance in statistical testing?

A. In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

Q. How to identify a cause vs. a correlation? Give examples?

A. While causation and correlation can exist at the same time, correlation does not imply causation. Causation explicitly applies to cases where action A causes outcome B. On the other hand, correlation is simply a relationship. Correlation between Ice

cream sales and sunglasses sold. As the sales of ice creams is increasing so do the sales of sunglasses. Causation takes a step further than correlation.

Q. what is precision, accuracy and recall?

A. The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. The precision is the proportion of relevant results in the list of all returned search results. Accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

Q. choose 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. word2vec methods?

A. Word2vec is a technique for natural language processing published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence.

Q. pruning in case of decision trees?

A. Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances.

Q. bag-of-words?

A. A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a “bag” of words because any information about the order or structure of words in the document is discarded.

Q. learning rate in gradient descent?

A. The learning rate is a configurable hyperparameter used in the training of neural networks that has a small positive value, often in the range between 0.0 and 1.0. The learning rate controls how quickly the model is adapted to the problem.

Q. What is Multicollinearity?

A. Multicollinearity occurs when two or more independent variables (also known as predictor) are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

Q. What's the difference between L1 and L2 regularization?

A. The main intuitive difference between the L1 and L2 regularization is that L1 regularization tries to estimate the median of the data while the L2 regularization tries to estimate the mean of the data to avoid overfitting. That value will also be the median of the data distribution mathematical

Q. Z score for outliers' treatment?

A. Z score test is one of the most commonly used methods to detect outliers. It measures the number of standard deviations away the observation is from the mean value. A z score of 1.5 indicated that the observation is 1.5 standard deviations above the mean and -1.5 means that the observation is 1.5 standard deviations below or less than the mean.

Q. Left and right Skweness?

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. Explain Object detection?

A. Object detection is a computer vision technique for locating instances of objects in images or videos. Object detection algorithms typically leverage machine learning or deep learning to produce meaningful results.
