Company Name - Accenture
Role - Data Scientist

Q. Difference between R square and Adjusted R Square?

A. One main difference between R^2 and the adjusted R^2: R^2 assumes that every single variable explains the variation in the dependent variable. The adjusted R^2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

Q. Difference between Precision and Recall?

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. Describe Assumptions of Linear Regression?

A. There are four assumptions associated with a linear regression model: Linearity: The relationship between X and the mean of Y is linear. Homoscedasticity: The variance of residual is the same for any value of X. Independence: Observations are independent of each other. The fourth one is normality.

Q. Difference between Random Forest and Decision Tree?

A. A decision tree combines some decisions, whereas a random forest combines several decision trees. Thus, it is a long process, yet slow. Whereas, a decision tree is fast and operates easily on large data sets, especially the linear one. The random forest model needs rigorous training.

Q. How does K-means work?

A. K-means clustering uses “centroids”, K different randomly-initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all of the points assigned to it.

Q. How do you generally choose among different classification models to decide which one is performing the best?

A. Here are some important considerations while choosing an algorithm:
Size of the training data, Accuracy and/or Interpretability of the output, Speed or Training time, Linearity and number of features.

Q. How do you perform feature selection?

A. Unsupervised: Do not use the target variable (e.g. remove redundant variables). Correlation.
Supervised: Use the target variable (e.g. remove irrelevant variables). Wrapper: Search for well-performing subsets of features. RFE.

Q. What is an intercept in a Linear Regression? What is its significance?

A. The intercept (often labelled as constant) is the point where the function crosses the y-axis. In some analysis, the regression model only becomes significant when we remove the intercept, and the regression line reduces to Y = b*X + error. The intercept (often labelled the constant) is the expected mean value of Y when all X="0. Start with a regression equation with one predictor, X. If X sometimes equals 0, the intercept is simply the expected mean value of Y at that value. If X never equals 0, then the intercept has no intrinsic meaning.

Q. Feature selection methods for selecting the right variables for building efficient predictive models?

A. Some of the Feature selection techniques are: Information Gain, Chi-square test, Correlation Coefficient, Mean Absolute Difference (MAD), Exhaustive selection, Forward selection, Regularization.

Q. Treat missing values?

A. They are:

- List wise or case deletion
- Pairwise deletion
- Mean substitution
- Regression imputation
- Maximum likelihood.

Q. assumptions used in linear regression? What would happen if they are violated?

A. The regression has five key assumptions:

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

Data to be analyzed by linear regression were sampled violate one or more of the linear regression assumptions, the results of the analysis may be incorrect or misleading.

Q. How is the grid search parameter different from the random search tuning strategy?

A. Random search differs from grid search in that we no longer provide an explicit set of possible values for each hyperparameter; rather, we provide a statistical distribution for each hyperparameter from which values are sampled. Essentially, we define a sampling distribution for each hyperparameter to carry out a randomized search.

Q. Is it good to do dimensionality reduction before fitting a Support Vector Model?

A. Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

Q. Imputation methods?

A. They are:

- List wise or case deletion
- Pairwise deletion
- Mean substitution
- Regression imputation
- Maximum likelihood.

Q. Model Pipeline? Benefits?

A. Automating the machine learning workflow by enabling data to be transformed and correlated into a model that can then be analysed to achieve outputs is done with ML pipelines. This type of ML pipeline makes the process of inputting data into the ML model fully automated. The main objective of having a proper pipeline for any ML model is to exercise control over it. A well-organised pipeline makes the implementation more flexible.

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