By OnlineInterviewQuestions.com

Data Scientist interview questions

A data scientist is an individual who is responsible for collecting, analyzing and interpreting large information regarding data to identify techniques. It will help a business to improve operations and reach greater heights in comparison to the competitors in the market. The ultimate role of a data scientist is to troubleshoot problems in different areas such as machine learning, predictive modelling and also provide visions and understandings beyond statistical analysis.

Some of the basic programming languages preferred by a data scientist are Python, R-Programming, SQL coding, Hand-loop platform, etc. A number of multinational companies these days are looking for individuals to help them grow in their business. Thus, such companies ask a variety of data scientist interview questions to not only freshers but also experienced individuals wishing to showcase their talent and knowledge in this field. Here are some important **Data scientist interview questions** that will not only give you a basic idea of the field but also help to clear the interview.

Q1. What is the difference between a cluster and systematic sampling?

Cluster Sampling is a technique that is used when studying a target population becomes difficult, especially a population spread across a wide area. While Systematic Sampling is a statistical technique where the list proceeds in a circular mode so that when one reaches the bottom of the list, it can be re-progressed back to the top.

Q2. How can you assess a good logistic model?

In order to assess a good logistic model, the following methods are employed:

- Using a classification metric to identify the correct negatives and incorrect positives
- Using concordance which helps to differentiate between the events that are going to happen and the ones that are not
- Using lift helps in comparing the logistic model with random selection

Q3. Describe the various steps involved while carrying out an analytical project.

The various steps carried out during an analytical project are:

- Understanding the business problem
- Exploring the data and familiarizing with it
- Preparing the data by spotting outliners, identifying and transforming variables and considering missing

values, etc.

- Once the data is prepared, start running the model, evaluate the result, make necessary changes in the approach (if required)
- Authenticate the model using a new set of data
- Start implementation of the model along with keeping a check on the result in order to evaluate the performance of the model over time

Q4. Why is dimensional reduction performed before fitting a Support Vector Machine (SVM)?

The reason for performing dimensional reduction before fitting an SVM is that it is best worked in a reduced space.

Q5. What is selection bias?

Selection bias is an active state when the sample data that is gathered and prepared has been characterized for modeling. However, it does not represent the true or future population that the model has to see.

Q6. What are Feature vectors?

Feature vectors are a type of n-dimensional vector that has various numerical features. They represent some item or a characteristic object. In the field of machine learning, feature vectors are important parameters that are used to represent different numeric or symbolic characteristics also known as features that represent an object in a mathematical way and can be easily analyzed.

Q7. What is the importance of selection bias?

Selection bias takes place when there is no suitable randomization obtained while selecting individuals, groups or data that has to be investigated. Selection bias simply indicates that the obtained sample does not exactly characterize the population that was essentially projected for analysis.

Q8. Differentiate between skewed and uniform distribution

Uniform distribution refers to a condition when all the observations in a dataset are equally spread across the range of distribution. Skewed distribution refers to the condition when one side of the graph has more dataset in comparison to the other side.

Q9. What is Machine Learning and how can it be used for time series analysis?

In simple terms, Machine Learning is the process when both the data and the equation is fed to the machine and it is directed to look into the data and identify the coefficient values in that equation. Yes, Machine Learning can be used for time series analysis.

Q10. Differentiate between Type I and Type II error

Type I error takes place when the null hypothesis is true; however, it is rejected.

Type II error occurs when the null hypothesis is false, but it is accepted as true.

Q11. Describe some assumptions considered important for linear regression.

Some of the assumptions that are considered important for linear regression are:

- There exists a linear relationship between the repressors and the dependent variables.
- The errors within the data need to be normally distributed and independent of each other.
- There should be a minimal multi-collinearity among the variables.

Q12. Give an example of a data set that has a non-Gaussian distribution

An example of a non-Gaussian distribution data is that of an exponential family of distributions in which there are more members with relevant skill set to be utilized in a varied field whenever necessary.

Q13. What is the purpose of A/B Testing?

A/B Testing is a statistical hypothesis for testing random experiment with two different variables A and B. The purpose of A/B testing is to categorize any changes that occur in the web pages to maximize or increase the outcome.

Q14. What is Difference between overfitting and underfitting?

Overfitting is a factual model that depicts irregular mistake or noise rather than the hidden relationship among variables. Overfitting happens when a model is unnecessarily unpredictable, for instance, when having a large number of parameters in respect to the number of perceptions. A model that has been overfitted has poor prescient execution, as it goes overboard to minor changes in the preparation information.

Underfitting happens when a factual model or machine learning calculation cannot catch the basic pattern of the information. Underfitting would happen, for instance, when fitting a direct model to non-straight information. Such a model also would have poor prescient execution.

Q15. Among Python and R, which one is generally preferred for text analytics?

- Python is generally preferred for text analytics because of the following reasons:
- Python has a Pandas library, which provides ease for usage of data structures
- Python performs faster for all kinds of text analytics

Q16. What is the importance of data cleaning in analysis?

The importance of data cleaning in the analysis are:

- Data cleaning from different sources helps in transforming data to a format that data scientist can use
- Cleaning of data can help in maximizing the accuracy of the model in machine learning
- Data cleaning is, however, a bulky procedure on the grounds that as the number of information sources builds, the time taken to clean the information increments exponentially because of the number of sources and the volume of information produced by these sources.
- 80% of the ideal opportunity might be simply used for cleaning the information that makes it a basic piece of investigation assignment.