Data Science Interview Questions for Freshers

1. What is Data Science?

Data Science is the area of study which involves extracting insights from vast amounts of data using various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data. The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.

2. What is the Difference Between Data Science and Machine Learning?

Data Science is a combination of algorithms, tools, and machine learning technique which helps you to find common hidden patterns from the given raw data. Whereas Machine learning is a branch of computer science, that deals with system programming to automatically learn and improve with experience.

3. Name three types of biases that can occur during sampling

In the sampling process, there are three types of biases, which are:

- Selection bias
- Under coverage bias
- Survivorship bias

4. Discuss Decision Tree algorithm

A decision tree is a popular supervised machine learning algorithm. It is mainly used for Regression and Classification. It allows breaks down a dataset into smaller subsets. The decision tree can able to handle both categorical and numerical data.

5. What is Prior probability and likelihood?

Prior probability is the proportion of the dependent variable in the data set while the likelihood is the probability of classifying a given observant in the presence of some other variable.

6. Explain Recommender Systems?

It is a subclass of information filtering techniques. It helps you to predict the preferences or ratings which users likely to give to a product.

7. Name three disadvantages of using a linear model

Three disadvantages of the linear model are:

- The assumption of linearity of the errors.
- You can't use this model for binary or count outcomes
- There are plenty of overfitting problems that it can't solve

8. Why do you need to perform resampling?

Resampling is done in below-given cases:

- Estimating the accuracy of sample statistics by drawing randomly with replacement from a set of the data point or using as subsets of accessible data
- Substituting labels on data points when performing necessary tests
- Validating models by using random subsets

9. List out the libraries in Python used for Data Analysis and Scientific Computations.

- SciPy
- Pandas
- Matplotlib
- NumPy
- SciKit
- Seaborn

10. What is Power Analysis?

The power analysis is an integral part of the experimental design. It helps you to determine the sample size requires to find out the effect of a given size from a cause with a specific level of assurance. It also allows you to deploy a particular probability in a sample size constraint.

11. Explain Collaborative filtering

Collaborative filtering used to search for correct patterns by collaborating viewpoints, multiple data sources, and various agents.

12. What is bias?

Bias is an error introduced in your model because of the oversimplification of a machine learning algorithm." It can lead to underfitting.

13. Discuss 'Naive' in a Naive Bayes algorithm?

The Naive Bayes Algorithm model is based on the Bayes Theorem. It describes the probability of an event. It is based on prior knowledge of conditions which might be related to that specific event.

14. What is a Linear Regression?

Linear regression is a statistical programming method where the score of a variable 'A' is predicted from the score of a second variable 'B'. B is referred to as the predictor variable and A as the criterion variable.

15. State the difference between the expected value and mean value

They are not many differences, but both of these terms are used in different contexts. Mean value is generally referred to when you are discussing a probability distribution whereas expected value is referred to in the context of a random variable.

16. What the aim of conducting A/B Testing?

AB testing used to conduct random experiments with two variables, A and B. The goal of this testing method is to find out changes to a web page to maximize or increase the outcome of a strategy.

17. What is Ensemble Learning?

The ensemble is a method of combining a diverse set of learners together to improvise on the stability and predictive power of the model. Two types of Ensemble learning methods are:

Bagging

Bagging method helps you to implement similar learners on small sample populations. It helps you to make nearer predictions.

Boosting

Boosting is an iterative method which allows you to adjust the weight of an observation depends upon the last classification. Boosting decreases the bias error and helps you to build strong predictive models.

18. Explain Eigenvalue and Eigenvector

Eigenvectors are for understanding linear transformations. Data scientist need to calculate the eigenvectors for a covariance matrix or correlation. Eigenvalues are the directions along using specific linear transformation acts by compressing, flipping, or stretching.

19. Define the term cross-validation

Cross-validation is a validation technique for evaluating how the outcomes of statistical analysis will generalize for an Independent dataset. This method is used in backgrounds where the objective is forecast, and one needs to estimate how accurately a model will accomplish.

20. Explain the steps for a Data analytics project

The following are important steps involved in an analytics project:

- Understand the Business problem
- Explore the data and study it carefully.
- Prepare the data for modeling by finding missing values and transforming variables.
- Start running the model and analyze the Big data result.
- Validate the model with new data set.
- Implement the model and track the result to analyze the performance of the model for a specific period.

21. Discuss Artificial Neural Networks

Artificial Neural networks (ANN) are a special set of algorithms that have revolutionized machine learning. It helps you to adapt according to changing input. So the network generates the best possible result without redesigning the output criteria.

22. What is Back Propagation?

Back-propagation is the essence of neural net training. It is the method of tuning the weights of a neural net depend upon the error rate obtained in

the previous epoch. Proper tuning of the helps you to reduce error rates and to make the model reliable by increasing its generalization.

23. What is a Random Forest?

Random forest is a machine learning method which helps you to perform all types of regression and classification tasks. It is also used for treating missing values and outlier values.

24. What is the importance of having a selection bias?

Selection Bias occurs when there is no specific randomization achieved while picking individuals or groups or data to be analyzed. It suggests that the given sample does not exactly represent the population which was intended to be analyzed.

25. What is the K-means clustering method?

K-means clustering is an important unsupervised learning method. It is the technique of classifying data using a certain set of clusters which is called K clusters. It is deployed for grouping to find out the similarity in the data.

Data Scientist Interview Questions for Experienced

26. Explain the difference between Data Science and Data Analytics

Data Scientists need to slice data to extract valuable insights that a data analyst can apply to real-world business scenarios. The main difference between the two is that the data scientists have more technical knowledge then business analyst. Moreover, they don’t need an understanding of the business required for data visualization.

27. Explain p-value?

When you conduct a hypothesis test in statistics, a p-value allows you to determine the strength of your results. It is a numerical number between 0 and 1. Based on the value it will help you to denote the strength of the specific result.

28. Define the term deep learning

Deep Learning is a subtype of machine learning. It is concerned with algorithms inspired by the structure called artificial neural networks (ANN).

29. Explain the method to collect and analyze data to use social media to predict the weather condition.

You can collect social media data using Facebook, twitter, Instagram's API's. For example, for the tweeter, we can construct a feature from each tweet like tweeted date, retweets, list of follower, etc. Then you can use a multivariate time series model to predict the weather condition.

RELATED ARTICLES

SAS Tutorial for Beginners: What is & Programming Example

Difference Between Data Science and Machine Learning

17 BEST Data Science Books (2025 Update)

Data Science Tutorial

30. When do you need to update the algorithm in Data science?

You need to update an algorithm in the following situation:

- You want your data model to evolve as data streams using infrastructure
- The underlying data source is changingIf it is non-stationarity

31. What is Normal Distribution

A normal distribution is a set of a continuous variable spread across a normal curve or in the shape of a bell curve. You can consider it as a continuous probability distribution which is useful in statistics. It is useful to analyze the variables and their relationships when we are using the normal distribution curve.

32. Which language is best for text analytics? R or Python?

Python will more suitable for text analytics as it consists of a rich library known as pandas. It allows you to use high-level data analysis tools and data structures, while R doesn't offer this feature.

33. Explain the benefits of using statistics by Data Scientists

Statistics help Data scientist to get a better idea of customer's expectation. Using the statistic method Data Scientists can get knowledge regarding consumer interest, behavior, engagement, retention, etc. It also helps you to build powerful data models to validate certain inferences and predictions.

34. Name various types of Deep Learning Frameworks

- Pytorch
- Microsoft Cognitive Toolkit
- TensorFlow
- Caffe
- Chainer
- Keras

35.Explain Auto-Encoder

Autoencoders are learning networks. It helps you to transform inputs into outputs with fewer numbers of errors. This means that you will get output to be as close to input as possible.

36. Define Boltzmann Machine

Boltzmann machines is a simple learning algorithm. It helps you to discover those features that represent complex regularities in the training data. This algorithm allows you to optimize the weights and the quantity for the given problem.

37. Explain why Data Cleansing is essential and which method you use to maintain clean data

Dirty data often leads to the incorrect inside, which can damage the prospect of any organization. For example, if you want to run a targeted marketing campaign. However, our data incorrectly tell you that a specific

product will be in-demand with your target audience; the campaign will fail.

38. What is skewed Distribution & uniform distribution?

Skewed distribution occurs when if data is distributed on any one side of the plot whereas uniform distribution is identified when the data is spread is equal in the range.

39. When underfitting occurs in a static model?

Underfitting occurs when a statistical model or machine learning algorithm not able to capture the underlying trend of the data.

40. What is reinforcement learning?

Reinforcement Learning is a learning mechanism about how to map situations to actions. The end result should help you to increase the binary reward signal. In this method, a learner is not told which action to take but instead must discover which action offers a maximum reward. As this method based on the reward/penalty mechanism.

41. Name commonly used algorithms.

Four most commonly used algorithm by Data scientist are:

- Linear regression
- Logistic regression
- Random Forest
- KNN

42. What is precision?

Precision is the most commonly used error metric is n classification mechanism. Its range is from 0 to 1, where 1 represents 100%

43. What is a univariate analysis?

An analysis which is applied to none attribute at a time is known as univariate analysis. Boxplot is widely used, univariate model.

44. How do you overcome challenges to your findings?

In order, to overcome challenges of my finding one need to encourage discussion, Demonstrate leadership and respecting different options.

45. Explain cluster sampling technique in Data science

A cluster sampling method is used when it is challenging to study the target population spread across, and simple random sampling can't be applied.

46. State the difference between a Validation Set and a Test Set

A Validation set mostly considered as a part of the training set as it is used for parameter selection which helps you to avoid overfitting of the model being built.

While a Test Set is used for testing or evaluating the performance of a trained machine learning model.

47. Explain the term Binomial Probability Formula?

"The binomial distribution contains the probabilities of every possible success on N trials for independent events that have a probability of π of occurring."

48. What is a recall?

A recall is a ratio of the true positive rate against the actual positive rate. It ranges from 0 to 1.

49. Discuss normal distribution

Normal distribution equally distributed as such the mean, median and mode are equal.

50. While working on a data set, how can you select important variables? Explain

Following methods of variable selection you can use:

- Remove the correlated variables before selecting important variables
- Use linear regression and select variables which depend on that p values.

- Use Backward, Forward Selection, and Stepwise Selection
- Use Xgboost, Random Forest, and plot variable importance chart.
- Measure information gain for the given set of features and select top n features accordingly.

51. Is it possible to capture the correlation between continuous and categorical variable?

Yes, we can use analysis of covariance technique to capture the association between continuous and categorical variables.

52. Treating a categorical variable as a continuous variable would result in a better predictive model?

Yes, the categorical value should be considered as a continuous variable only when the variable is ordinal in nature. So it is a better predictive model.