**Data Science Interview Questions**

**1. Compare Data Science Vs. Machine Learning**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Data Science** | **Machine Learning** |
| Scope | Multidisciplinary | Training machines |
| Artificial Intelligence | Loosely integrated | Tightly integrated |
| Role | Can take on a business role | Purely technical role |

**2. What is Data Science?**

Data Science is a blend of Statistics, technical skills and business vision which is used to analyze the available data and predict the future trend.

**3. How is it different from Big Data and Data Analytics?**

|  |  |  |
| --- | --- | --- |
| **Big Data** | **Data Science** | **Data Analytics** |
| Huge volumes of data-structured, unstructured and semi-structured | Deals with slicing and dicing the data | Contributing operational insights into complex business scenarios |
| Requires a basic knowledge of statistics and mathematics | Requires in-depth knowledge of statistics and mathematics | Requires moderate amount of statistics and mathematics |

**4. Which language is more suitable for text analytics? R or Python?**

Since Python consists of a rich library called Pandas which allows the analysts to use high-level data analysis tools as well as data structures, while R lacks this feature. Hence Python will more suitable for text analytics.

**5. What is a Recommender System?**

A recommender system is today widely deployed in multiple fields like movie recommendations, music preferences, social tags, research articles, search queries and so on. The recommender systems work as per collaborative and content-based filtering or by deploying a personality-based approach. This type of system works based on a person’s past behavior in order to build a model for the future. This will predict the future product buying, movie viewing or book reading by people. It also creates a filtering approach using the discrete characteristics of items while recommending additional items.

**6. Compare SAS, R and Python programming?**

SAS: it is one of the most widely used analytics tools used by some of the biggest companies on earth. It has some of the best statistical functions, graphical user interface, but can come with a price tag and hence it cannot be readily adopted by smaller enterprises  
R: The best part about R is that it is an Open Source tool and hence used generously by academia and the research community. It is a robust tool for statistical computation, graphical representation and reporting. Due to its open source nature it is always being updated with the latest features and then readily available to everybody.  
Python: Python is a powerful open source programming language that is easy to learn, works well with most other tools and technologies. The best part about Python is that it has innumerable libraries and community created modules making it very robust. It has functions for statistical operation, model building and more.

**7. Explain the various benefits of R language?**

The R programming language includes a set of software suite that is used for graphical representation, statistical computing, data manipulation and calculation.  
Some of the highlights of R programming environment include the following:

* An extensive collection of tools for data analysis
* Operators for performing calculations on matrix and array
* Data analysis technique for graphical representation
* A highly developed yet simple and effective programming language
* It extensively supports machine learning applications
* It acts as a connecting link between various software, tools and datasets
* Create high quality reproducible analysis that is flexible and powerful
* Provides a robust package ecosystem for diverse needs
* It is useful when you have to solve a data-oriented problem

**8. What are the two main components of the Hadoop Framework?**

HDFS and YARN are basically the two major components of Hadoop framework.

* **HDFS-** Stands for Hadoop Distributed File System. It is the distributed database working on top of Hadoop. It is capable of storing and retrieving bulk of datasets in no time.
* **YARN-** Stands for Yet Another Resource Negotiator. It allocates resources dynamically and handles the workloads.

**9. How do Data Scientists use Statistics?**

Statistics helps Data Scientists to look into the data for patterns, hidden insights and convert Big Data into Big insights. It helps to get a better idea of what the customers are expecting. Data Scientists can learn about the consumer behavior, interest, engagement, retention and finally conversion all through the power of insightful statistics. It helps them to build powerful data models in order to validate certain inferences and predictions. All this can be converted into a powerful business proposition by giving users what they want at precisely when they want it.

**10. What is logistic regression?**

It is a statistical technique or a model in order to analyze a dataset and predict the binary outcome. The outcome has to be a binary outcome that is either zero or one or a yes or no. Random Forest is an important technique which is used to do classification, regression and other tasks on data.

**11. Why data cleansing is important in data analysis?**

With data coming in from multiple sources it is important to ensure that data is good enough for analysis. This is where data cleansing becomes extremely vital. Data cleansing extensively deals with the process of detecting and correcting of data records, ensuring that data is complete and accurate and the components of data that are irrelevant are deleted or modified as per the needs. This process can be deployed in concurrence with data wrangling or batch processing.  
Once the data is cleaned it confirms with the rules of the data sets in the system. Data cleansing is an essential part of the data science because the data can be prone to error due to human negligence, corruption during transmission or storage among other things. Data cleansing takes a huge chunk of time and effort of a Data Scientist because of the multiple sources from which data emanates and the speed at which it comes.

**12. Describe univariate, bivariate and multivariate analysis.**

As the name suggests these are analysis methodologies having a single, double or multiple variables.  
So a univariate analysis will have one variable and due to this there are no relationships, causes. The major aspect of the univariate analysis is to summarize the data and find the patterns within it to make actionable decisions.  
A Bivariate analysis deals with the relationship between two sets of data. These sets of paired data come from related sources, or samples. There are various tools to analyze such data including the chi-squared tests and t-tests when the data are having a correlation. If the data can be quantified then it can analyzed using a graph plot or a scatterplot. The strength of the correlation between the two data sets will be tested in a Bivariate analysis.

**13. How machine learning is deployed in real world scenarios?**

Here are some of the scenarios in which machine learning finds applications in real world:

* Ecommerce: Understanding the customer churn, deploying targeted advertising, remarketing
* Search engine: Ranking pages depending on the personal preferences of the searcher
* Finance: Evaluating investment opportunities & risks, detecting fraudulent transactions
* Medicare: Designing drugs depending on the patient’s history and needs
* Robotics: Machine learning for handling situations that are out of the ordinary
* Social media: Understanding relationships and recommending connections
* Extraction of information: framing questions for getting answers from databases over the web

**14. What are the various aspects of a Machine Learning process?**

In this post I will discuss the components involved in solving a problem using machine learning.

* Domain knowledge  
  This is the first step wherein we need to understand how to extract the various features from the data and learn more about the data that we are dealing with. It has got more to do with the type of domain that we are dealing with and familiarizing the system to learn more about it.
* Feature Selection  
  This step has got more to do with the feature that we are selecting from the set of features that we have. Sometimes it happens that there are a lot of features and we have to make an intelligent decision regarding the type of feature that we want to select to go ahead with our machine learning endeavor.
* Algorithm  
  This is a vital step since the algorithms that we choose will have a very major impact on the entire process of machine learning. You can choose between the linear and nonlinear algorithm. Some of the algorithms used are Support Vector Machines, Decision Trees, Naïve Bayes, K-Means Clustering, etc.
* Training  
  This is the most important part of the machine learning technique and this is where it differs from the traditional programming. The training is done based on the data that we have and providing more real world experiences. With each consequent training step the machine gets better and smarter and able to take improved decisions.
* Evaluation  
  In this step we actually evaluate the decisions taken by the machine in order to decide whether it is up to the mark or not. There are various metrics that are involved in this process and we have to closed deploy each of these to decide on the efficacy of the whole machine learning endeavor.
* Optimization  
  This process involves improving the performance of the machine learning process using various optimization techniques. Optimization of machine learning is one of the most vital components wherein the performance of the algorithm is vastly improved. The best part of optimization techniques is that machine learning is not just a consumer of optimization techniques but it also provides new ideas for optimization too.
* Testing  
  Here various tests are carried out and some these are unseen set of test cases. The data is partitioned into test and training set. There are various testing techniques like cross-validation in order to deal with multiple situations.

**15. What do you understand by the term Normal Distribution?**

It is a set of continuous variable spread across a normal curve or in the shape of a bell curve. It can be considered as a continuous probability distribution and is useful in statistics. It is the most common distribution curve and it becomes very useful to analyze the variables and their relationships when we have the normal distribution curve.  
The normal distribution curve is symmetrical. The non-normal distribution approaches the normal distribution as the size of the samples increases. It is also very easy to deploy the Central Limit Theorem. This method helps to make sense of data that is random by creating an order and interpreting the results using a bell-shaped graph.

**16. What is Linear Regression?**

It is the most commonly used method for predictive analytics. The Linear Regression method is used to describe relationship between a dependent variable and one or independent variable. The main task in the Linear Regression is the method of fitting a single line within a scatter plot. The Linear Regression consists of the following three methods:

* Determining and analyzing the correlation and direction of the data
* Deploying the estimation of the model
* Ensuring the usefulness and validity of the model  
  It is extensively used in scenarios where the cause effect model comes into play. For example you want to know the effect of a certain action in order to determine the various outcomes and extent of effect the cause has in determining the final outcome.

**17. What is Interpolation and Extrapolation?**

The terms of interpolation and extrapolation are extremely important in any statistical analysis. Extrapolation is the determination or estimation using a known set of values or facts by extending it and taking it to an area or region that is unknown. It is the technique of inferring something using data that is available.  
Interpolation on the other hand is the method of determining a certain value which falls between a certain set of values or the sequence of values. This is especially useful when you have data at the two extremities of a certain region but you don’t have enough data points at the specific point. This is when you deploy interpolation to determine the value that you need.

**18. What is Power Analysis?**

The power analysis is a vital part of the experimental design. It is involved with the process of determining the sample size needed for detecting an effect of a given size from a cause with a certain degree of assurance. It lets you deploy specific probability in a sample size constraint.  
The various techniques of statistical power analysis and sample size estimation are widely deployed for making statistical judgment that are accurate and evaluate the size needed for experimental effects in practice.  
Power analysis lets you understand the sample size estimate so that they are neither high nor low. A low sample size there will be no authentication to provide reliable answers and if it is large there will be wastage of resources.

**19. What is K-means? How can you select K for K-means?**

K-means clustering can be termed as the basic unsupervised learning algorithm. It is the method of classifying data using a certain set of clusters called as K clusters. It is deployed for grouping data in order to find similarity in the data.  
It includes defining the K centers, one each in a cluster. The clusters are defined into K groups with K being predefined. The K points are selected at random as cluster centers. The objects are assigned to their nearest cluster center. The objects within a cluster are as closely related to one another as possible and differ as much as possible to the objects in other clusters. K-means clustering works very well for large sets of data.

**20. How is Data modeling different from Database design?**

Data Modeling: It can be considered as the first step towards the design of a database. Data modeling creates a conceptual model based on the relationship between various data models. The process involves moving from the conceptual stage to the logical model to the physical schema. It involves the systematic method of applying the data modeling techniques.  
Database Design: This is the process of designing the database. The database design creates an output which is a detailed data model of the database. Strictly speaking database design includes the detailed logical model of a database but it can also include physical design choices and storage parameters.