What is Data Science?

Data Science combines statistics, maths, specialised programs, artificial intelligence, machine learning etc. Data Science is simply the application of specific principles and analytic techniques to extract information from data used in strategic planning, decision making, etc. Simply, data science means analysing data for actionable insights.

Differentiate Between Data Analytics and Data Science

Data Analytics	*Data Science*
Data Analytics use data to draw meaningful insights and solves problems.	*Data Science is used in asking questions, writing algorithms, coding and building statistical models.*
Data analytics tools include data mining, data modelling, database management and data analysis.	*Machine Learning, Hadoop, Java, Python, software development etc., are the tools of Data Science.*
Use the existing information to uncover the actionable data.	*As a result, data Science discovers new Questions to drive innovation.*
Check data from the given information using a specialised system and software.	*This field uses scientific methods and algorithms to extract knowledge from unstructured data.*

Basic and Advanced Data Science Interview Questions

Here's a list of the most popular data science interview questions on the technical concept which you can expect to face, and how to frame your answers.

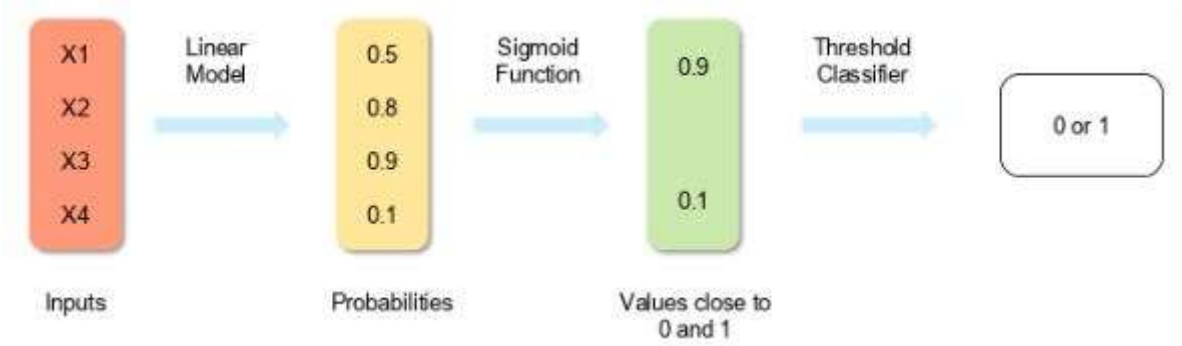
1. What are the differences between supervised and unsupervised learning?

Supervised Learning	*Unsupervised Learning*
• *Uses known and labeled data as input* • *Supervised learning has a feedback mechanism* • *The most commonly used supervised learning algorithms are decision trees, logistic regression, and support vector machine*	• *Uses unlabeled data as input* • *Unsupervised learning has no feedback mechanism* • *The most commonly used unsupervised learning algorithms are k-means clustering, hierarchical clustering, and apriori algorithm*

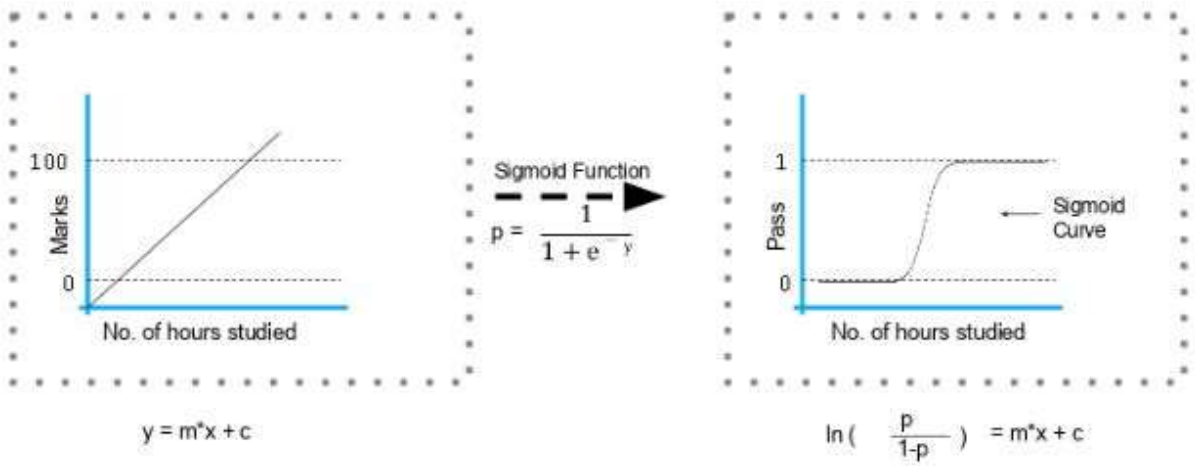
2. How is logistic regression done?

Logistic regression measures the relationship between the dependent variable (our label of what we want to predict) and one or more independent variables (our features) by estimating probability using its underlying logistic function (sigmoid).

The image shown below depicts how logistic regression works:



The formula and graph for the sigmoid function are as shown:



3. Explain the steps in making a decision tree.

1. *Take the entire data set as input*
2. *Calculate entropy of the target variable, as well as the predictor attributes*
3. *Calculate your information gain of all attributes (we gain information on sorting different objects from each other)*
4. *Choose the attribute with the highest information gain as the root node*
5. *Repeat the same procedure on every branch until the decision node of each branch is finalized*

For example, let's say you want to build a [decision tree] to decide whether you should accept or decline a job offer. The decision tree for this case is as shown:

It is clear from the decision tree that an offer is accepted if:

- *Salary is greater than $50,000*

- *The commute is less than an hour*
- *Incentives are offered*

4. How do you build a random forest model?

A random forest is built up of a number of decision trees. If you split the data into different packages and make a decision tree in each of the different groups of data, the random forest brings all those trees together.

Steps to build a random forest model:

1. *Randomly select 'k' features from a total of 'm' features where k << m*
2. *Among the 'k' features, calculate the node D using the best split point*
3. *Split the node into daughter nodes using the best split*
4. *Repeat steps two and three until leaf nodes are finalized*
5. *Build forest by repeating steps one to four for 'n' times to create 'n' number of trees*

5. How can you avoid overfitting your model?

Overfitting refers to a model that is only set for a very small amount of data and ignores the bigger picture. There are three main methods to avoid overfitting:

1. *Keep the model simple—take fewer variables into account, thereby removing some of the noise in the training data*
2. *Use cross-validation techniques, such as k folds cross-validation*
3. *Use regularization techniques, such as LASSO, that penalize certain model parameters if they're likely to cause overfitting*👊

6. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate

Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.

Example: height of students

Height (in cm)
164
167.3
170
174.2
178
180

The patterns can be studied by drawing conclusions using mean, median, mode, dispersion or range, minimum, maximum, etc.

Bivariate

Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.

Example: temperature and ice cream sales in the summer season

Temperature (in Celcius)	*Sales*
20	*2,000*
25	*2,100*

26	*2,300*
28	*2,400*
30	*2,600*
36	*3,100*

Here, the relationship is visible from the table that temperature and sales are directly proportional to each other. The hotter the temperature, the better the sales.

Multivariate

Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable.

Example: data for house price prediction

No. of rooms	*Floors*	*Area (sq ft)*	*Price*
2	*0*	*900*	*$4000,00*
3	*2*	*1,100*	*$600,000*
3.5	*5*	*1,500*	*$900,000*
4	*3*	*2,100*	*$1,200,000*

The patterns can be studied by drawing conclusions using mean, median, and mode, dispersion or range, minimum, maximum, etc. You can start describing the data and using it to guess what the price of the house will be.

7. What are the feature selection methods used to select the right variables?

There are two main methods for feature selection, i.e, filter, and wrapper methods.

Filter Methods

This involves:

- *Linear discrimination analysis*
- ***ANOVA***
- *Chi-Square*

The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting the features, it's all about cleaning up the data coming in.

Wrapper Methods

This involves:

- *Forward Selection: We test one feature at a time and keep adding them until we get a good fit*
- *Backward Selection: We test all the features and start removing them to see what works better*
- *Recursive Feature Elimination: Recursively looks through all the different features and how they pair together*

Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is performed with the wrapper method.

8. In your choice of language, write a program that prints the numbers ranging from one to 50.

But for multiples of three, print "Fizz" instead of the number, and for the multiples of five, print "Buzz." For numbers which are multiples of both three and five, print "FizzBuzz"

The code is shown below:

```
for fizzbuzz in range(51):
    if fizzbuzz % 3 == 0 and fizzbuzz % 5 == 0:
        print("fizzbuzz")
        continue
    elif fizzbuzz % 3 == 0:
        print("fizz")
        continue
    elif fizzbuzz % 5 == 0:
        print("buzz")
        continue
    print(fizzbuzz)
```

Note that the range mentioned is 51, which means zero to 50. However, the range asked in the question is one to 50. Therefore, in the above code, you can include the range as (1,51).

9. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

The following are ways to handle missing data values:

If the data set is large, we can just simply remove the rows with missing data values. It is the quickest way; we use the rest of the data to predict the values.

For smaller data sets, we can substitute missing values with the mean or average of the rest of the data using the pandas' data frame in python. There are different ways to do so, such as df.mean(), df.fillna(mean).

10. For the given points, how will you calculate the Euclidean distance in Python?

plot1 = [1,3]

plot2 = [2,5]

The Euclidean distance can be calculated as follows:

*euclidean_distance = sqrt((plot1[0]-plot2[0])**2 + (plot1[1]-plot2[1])**2)*

Check out the Simplilearn's video on "Data Science Interview Question" curated by industry experts to help you prepare for an interview.

11. What are dimensionality reduction and its benefits?

The Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.

This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

12. How will you calculate eigenvalues and eigenvectors of the following 3x3 matrix?

-2	*-4*	*2*
-2	*1*	*2*
4	*2*	*5*

The characteristic equation is as shown:

Expanding determinant:

(-2 – λ) [(1-λ) (5-λ)-2x2] + 4[(-2) x (5-λ) -4x2] + 2[(-2) x 2-4(1-λ)] =0

- λ3 + 4λ2 + 27λ – 90 = 0,

λ3 - 4 λ2 -27 λ + 90 = 0

Here we have an algebraic equation built from the eigenvectors.

By hit and trial:

33 – 4 x 32 - 27 x 3 +90 = 0

Hence, (λ - 3) is a factor:

λ3 - 4 λ2 - 27 λ +90 = (λ – 3) (λ2 – λ – 30)

Eigenvalues are 3,-5,6:

(λ – 3) (λ2 – λ – 30) = (λ – 3) (λ+5) (λ-6),

Calculate eigenvector for λ = 3

For X = 1,

-5 - 4Y + 2Z =0,

-2 - 2Y + 2Z =0

Subtracting the two equations:

3 + 2Y = 0,

Subtracting back into second equation:

Y = -(3/2)

Z = -(1/2)

Similarly, we can calculate the eigenvectors for -5 and 6.

13. How should you maintain a deployed model?

The steps to maintain a deployed model are:

Monitor

Constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it's doing what it's supposed to do.

Evaluate

Evaluation metrics of the current model are calculated to determine if a new algorithm is needed.

Compare

The new models are compared to each other to determine which model performs the best.

Rebuild

The best performing model is re-built on the current state of data.

14. What are recommender systems?

A recommender system predicts what a user would rate a specific product based on their preferences. It can be split into two different areas:

Collaborative Filtering

As an example, Last.fm recommends tracks that other users with similar interests play often. This is also commonly seen on Amazon after making a purchase; customers may notice the following message accompanied by product recommendations: "Users who bought this also bought…"

Content-based Filtering

As an example: Pandora uses the properties of a song to recommend music with similar properties. Here, we look at content, instead of looking at who else is listening to music.

15. How do you find RMSE and MSE in a linear regression model?

RMSE and MSE are two of the most common measures of accuracy for a linear regression model.

RMSE indicates the Root Mean Square Error.

```
> rmse
[1] 3.339665e-11
```

MSE indicates the Mean Square Error.

$$MSE = \frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}$$

16. How can you select k for k-means?

We use the elbow method to select k for k-means clustering. The idea of the elbow method is to run k-means clustering on the data set where 'k' is the number of clusters.

Within the sum of squares (WSS), it is defined as the sum of the squared distance between each member of the cluster and its centroid.

17. What is the significance of p-value?

p-value typically ≤ 0.05

This indicates strong evidence against the null hypothesis; so you reject the null hypothesis.

p-value typically > 0.05

This indicates weak evidence against the null hypothesis, so you accept the null hypothesis.

p-value at cutoff 0.05

This is considered to be marginal, meaning it could go either way.

18. How can outlier values be treated?

You can drop outliers only if it is a garbage value.

Example: height of an adult = abc ft. This cannot be true, as the height cannot be a string value. In this case, outliers can be removed.

If the outliers have extreme values, they can be removed. For example, if all the data points are clustered between zero to 10, but one point lies at 100, then we can remove this point.

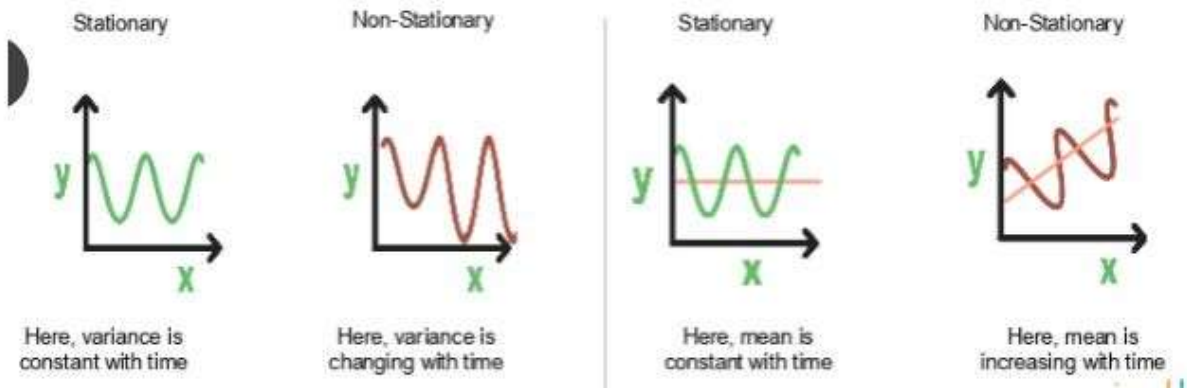
If you cannot drop outliers, you can try the following:

- *Try a different model. Data detected as outliers by linear models can be fit by nonlinear models. Therefore, be sure you are choosing the correct model.*
- *Try normalizing the data. This way, the extreme data points are pulled to a similar range.*
- *You can use algorithms that are less affected by outliers; an example would be random forests.*

19. How can time-series data be declared as stationery?

It is stationary when the variance and mean of the series are constant with time.

Here is a visual example:

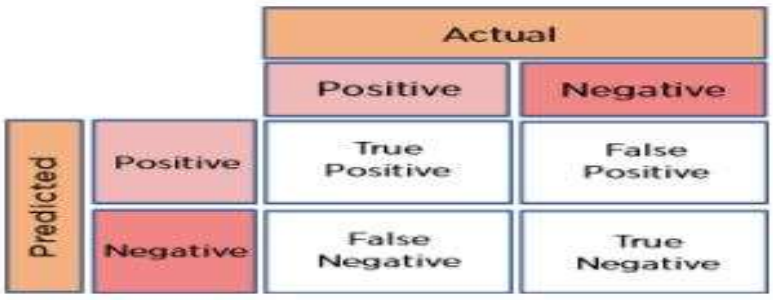


In the first graph, the variance is constant with time. Here, X is the time factor and Y is the variable. The value of Y goes through the same points all the time; in other words, it is stationary.

In the second graph, the waves get bigger, which means it is non-stationary and the variance is changing with time.

20. How can you calculate accuracy using a confusion matrix?

*Consider this [confusion matrix](): *



You can see the values for total data, actual values, and predicted values.

The formula for accuracy is:

Accuracy = (True Positive + True Negative) / Total Observations

= (262 + 347) / 650

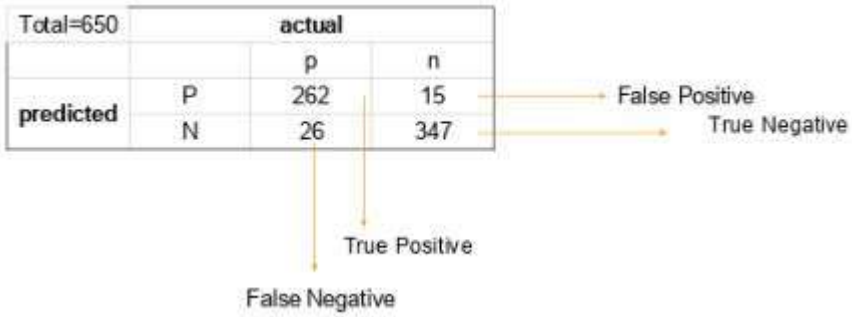
= 609 / 650

= 0.93

As a result, we get an accuracy of 93 percent.

21. Write the equation and calculate the precision and recall rate.

Consider the same confusion matrix used in the previous question.



Precision = (True positive) / (True Positive + False Positive)

= 262 / 277

= 0.94

Recall Rate = (True Positive) / (Total Positive + False Negative)

= 262 / 288

= 0.90

22. 'People who bought this also bought…' recommendations seen on Amazon are a result of which algorithm?

The recommendation engine is accomplished with collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.

The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.

For example, a sales page shows that a certain number of people buy a new phone and also buy tempered glass at the same time. Next time, when a person buys a phone, he or she may see a recommendation to buy tempered glass as well.

23. Write a basic SQL query that lists all orders with customer information.

Usually, we have order tables and customer tables that contain the following columns:

- *Order Table*
- *Orderid*
- *customerId*
- *OrderNumber*
- *TotalAmount*
- *Customer Table*
- *Id*
- *FirstName*
- *LastName*
- *City*
- *Country*
- *The SQL query is:*
- *SELECT OrderNumber, TotalAmount, FirstName, LastName, City, Country*
- *FROM Order*
- *JOIN Customer*
- *ON Order.CustomerId = Customer.Id*

24. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why

shouldn't you be happy with your model performance? What can you do about it?

Cancer detection results in imbalanced data. In an imbalanced dataset, accuracy should not be based as a measure of performance. It is important to focus on the remaining four percent, which represents the patients who were wrongly diagnosed. Early diagnosis is crucial when it comes to cancer detection, and can greatly improve a patient's prognosis.

Hence, to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine the class wise performance of the classifier.

25. Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables?

- *K-means clustering*
- *Linear regression*
- *K-NN (k-nearest neighbor)*
- *Decision trees*

The K nearest neighbor algorithm can be used because it can compute the nearest neighbor and if it doesn't have a value, it just computes the nearest neighbor based on all the other features.

When you're dealing with K-means clustering or linear regression, you need to do that in your pre-processing, otherwise, they'll crash. Decision trees also have the same problem, although there is some variance.

26. Below are the eight actual values of the target variable in the train file. What is the entropy of the target variable?

[0, 0, 0, 1, 1, 1, 1, 1]

Choose the correct answer.

1. *-(5/8 log(5/8) + 3/8 log(3/8))*
2. *5/8 log(5/8) + 3/8 log(3/8)*
3. *3/8 log(5/8) + 5/8 log(3/8)*
4. *5/8 log(3/8) – 3/8 log(5/8)*

The target variable, in this case, is 1.

The formula for calculating the entropy is:

Putting p=5 and n=8, we get

Entropy = A = -(5/8 log(5/8) + 3/8 log(3/8))

27. We want to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case?

Choose the correct option:

1. *Logistic Regression*
2. *Linear Regression*
3. *K-means clustering*
4. *Apriori algorithm*

The most appropriate algorithm for this case is A, logistic regression.

28. After studying the behavior of a population, you have identified four specific individual types that are valuable to your study. You would like to find all users who are most similar to each individual type. Which algorithm is most appropriate for this study?

Choose the correct option:

1. *K-means clustering*
2. *Linear regression*
3. *Association rules*
4. *Decision trees*

As we are looking for grouping people together specifically by four different similarities, it indicates the value of k. Therefore, K-means clustering (answer A) is the most appropriate algorithm for this study.

29. You have run the association rules algorithm on your dataset, and the two rules {banana, apple} => {grape} and {apple, orange} => {grape} have been found to be relevant. What else must be true?

Choose the right answer:

1. *{banana, apple, grape, orange} must be a frequent itemset*
2. *{banana, apple} => {orange} must be a relevant rule*
3. *{grape} => {banana, apple} must be a relevant rule*
4. *{grape, apple} must be a frequent itemset*

The answer is A: {grape, apple} must be a frequent itemset

30. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should you use?

1. *One-way ANOVA*
2. *K-means clustering*
3. *Association rules*
4. *Student's t-test*

The answer is A: One-way ANOVA

31. What do you understand about true positive rate and false-positive rate?

- *The True Positive Rate (TPR) defines the probability that an actual positive will turn out to be positive.*

The True Positive Rate (TPR) is calculated by taking the ratio of the [True Positives (TP)] and [True Positive (TP) & False Negatives (FN)].

The formula for the same is stated below -

TPR=TP/TP+FN

- *The False Positive Rate (FPR) defines the probability that an actual negative result will be shown as a positive one i.e the probability that a model will generate a false alarm.*

The False Positive Rate (FPR) is calculated by taking the ratio of the [False Positives (FP)] and [True Positives (TP) & False Positives(FP)].

The formula for the same is stated below -

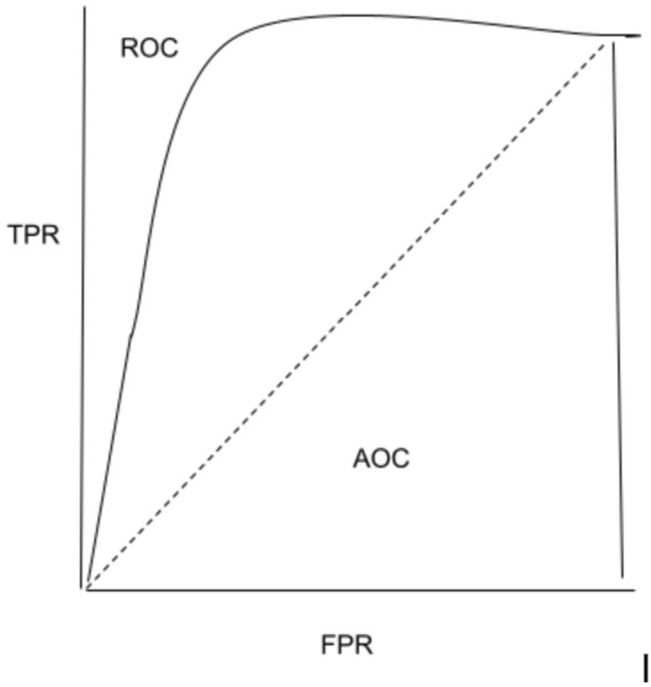
FPR=FP/TN+FP

32. What is the ROC curve?

The graph between the True Positive Rate on the y-axis and the False Positive Rate on the x-axis is called the ROC curve and is used in binary classification.

The False Positive Rate (FPR) is calculated by taking the ratio between False Positives and the total number of negative samples, and the True Positive Rate (TPR) is calculated by taking the ratio between True Positives and the total number of positive samples.

In order to construct the ROC curve, the TPR and FPR values are plotted on multiple threshold values. The area range under the ROC curve has a range between 0 and 1. A completely random model, which is represented by a straight line, has a 0.5 ROC. The amount of deviation a ROC has from this straight line denotes the efficiency of the model.



The image above denotes a ROC curve example.

33. What is a Confusion Matrix?

*The Confusion Matrix is the summary of prediction results of a particular problem. It is a table that is used to describe the performance of the model. The Confusion Matrix is an n*n matrix that evaluates the performance of the classification model.*

34. What do you understand about the true-positive rate and false-positive rate?

TRUE-POSITIVE RATE: The true-positive rate gives the proportion of correct predictions of the positive class. It is also used to measure the percentage of actual positives that are accurately verified.

FALSE-POSITIVE RATE: The false-positive rate gives the proportion of incorrect predictions of the positive class. A false positive determines something is true when that is initially false.

35. How is Data Science different from traditional application programming?

The primary and vital difference between Data Science and traditional application programming is that in traditional programming, one has to create rules to translate the input to output. In Data Science, the rules are automatically produced from the data.

36. What is the difference between the long format data and wide format data?

LONG FORMAT DATA: It contains values that repeat in the first column. In this format, each row is a one-time point per subject.

WIDE FORMAT DATA: In the Wide Format Data, the data's repeated responses will be in a single row, and each response can be recorded in separate columns.

Long format Table:

NAME	*ATTRIBUTE*	*VALUE*

RAMA	*HEIGHT*	*182*
SITA	*HEIGHT*	*160*

Wide format Table:

NAME	*HEIGHT*
RAMA	*182*
SITA	*160*

37. Mention some techniques used for sampling. What is the main advantage of sampling?

Sampling is the selection of individual members or a subset of the population to estimate the characters of the whole population. There are two types of Sampling, namely Probability and Non-Probability Sampling.

38. Why is Python used for Data Cleaning in DS?

Data Scientists and technical analysts must convert a huge amount of data into effective ones. Data Cleaning includes removing malwared records, outliners, inconsistent values, redundant formatting etc. Matplotlib, Pandas etc are the most used Python Data Cleaners.

39. What are the popular libraries used in Data Science?

The popular libraries used in Data Science are

- ***Tensor Flow***
- ***Pandas***
- ***NumPy***
- ***SciPy***

- *Scrapy*
- *Librosa*
- *MatPlotLib*

40. What is variance in Data Science?

Variance is the value which depicts the individual figures in a set of data which distributes themselves about the mean and describes the difference of each value from the mean value. Data Scientists use variance to understand the distribution of a data set.

41. What is pruning in a decision tree algorithm?

In Data Science and Machine Learning, Pruning is a technique which is related to decision trees. Pruning simplifies the decision tree by reducing the rules. Pruning helps to avoid complexity and improves accuracy. Reduced error Pruning, cost complexity pruning etc. are the different types of Pruning.

42. What is entropy in a decision tree algorithm?

Entropy is the measure of randomness or disorder in the group of observations. It also determines how a decision tree switches to split data. Entropy is also used to check the homogeneity of the given data. If the entropy is zero, then the sample of data is entirely homogeneous, and if the entropy is one, then it indicates that the sample is equally divided.

43. What information is gained in a decision tree algorithm?

Information gain is the expected reduction in entropy. Information gain decides the building of the tree. Information Gain makes the decision tree smarter. Information gain includes parent node R and a set E of K training examples. It calculates the difference between entropy before and after the split.

44. What is k-fold cross-validation?

The k-fold cross validation is a procedure used to estimate the model's skill in new data. In k-fold cross validation, every observation from the original dataset may appear in the training and testing set. K-fold cross-validation estimates the accuracy but does not help you to improve the accuracy.

45. What is a normal distribution?

Normal Distribution is also known as the Gaussian Distribution. The normal distribution shows the data near the mean and the frequency of that particular data. When represented in graphical form, normal distribution appears like a bell curve. The parameters included in the normal distribution are Mean, Standard Deviation, Median etc.

46. What is Deep Learning?

Deep Learning is one of the essential factors in Data Science, including statistics. Deep Learning makes us work more closely with the human brain and reliable with human thoughts. The algorithms are sincerely created to resemble the human brain. In Deep Learning, multiple layers are formed from the raw input to extract the high-level layer with the best features.

47. What is an RNN (recurrent neural network)?

RNN is an algorithm that uses sequential data. RNN is used in language translation, voice recognition, image capturing etc. There are different types of RNN networks such as one-to-one, one-to-many, many-to-one and many-to-many. RNN is used in Google's Voice search and Apple's Siri.

Basic Data Science Interview Questions

Let us begin with a few basic data science interview questions!

48. What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an object in a mathematical way that's easy to analyze.

49. What are the steps in making a decision tree?

1. *Take the entire data set as input.*
2. *Look for a split that maximizes the separation of the classes. A split is any test that divides the data into two sets.*
3. *Apply the split to the input data (divide step).*
4. *Re-apply steps one and two to the divided data.*
5. *Stop when you meet any stopping criteria.*

6. *This step is called pruning. Clean up the tree if you went too far doing splits.*

50. What is root cause analysis?

Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring.

51. What is logistic regression?

Logistic regression is also known as the logit model. It is a technique used to forecast the binary outcome from a linear combination of predictor variables.

52. What are recommender systems?

Recommender systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product.

53. Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. It is mainly used in backgrounds where the objective is to forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) to limit problems like overfitting and gain insight into how the model will generalize to an independent data set.

54. What is collaborative filtering?

Most recommender systems use this filtering process to find patterns and information by collaborating perspectives, numerous data sources, and several agents.

55. Do gradient descent methods always converge to similar points?

They do not, because in some cases, they reach a local minima or a local optima point. You would not reach the global optima point. This is governed by the data and the starting conditions.

56. What is the goal of A/B Testing?

This is statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B testing is to detect any changes to a web page to maximize or increase the outcome of a strategy.

57. What are the drawbacks of the linear model?

- *The assumption of linearity of the errors*
- *It can't be used for count outcomes or binary outcomes*
- *There are overfitting problems that it can't solve*

58. What is the law of large numbers?

It is a theorem that describes the result of performing the same experiment very frequently. This theorem forms the basis of frequency-style thinking. It states that the sample mean, sample variance, and sample standard deviation converge to what they are trying to estimate.

59. What are the confounding variables?

These are extraneous variables in a statistical model that correlates directly or inversely with both the dependent and the independent variable. The estimate fails to account for the confounding factor.

60. What is star schema?

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes, star schemas involve several layers of summarization to recover information faster.

61. How regularly must an algorithm be updated?

You will want to update an algorithm when:

- *You want the model to evolve as data streams through infrastructure*
- *The underlying data source is changing*
- *There is a case of non-stationarity*

62. What are eigenvalue and eigenvector?

Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing, or stretching.

Eigenvectors are for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix.

63. Why is resampling done?

Resampling is done in any of these cases:

- ***Estimating the accuracy of sample statistics by using subsets of accessible data, or drawing randomly with replacement from a set of data points***
- ***Substituting labels on data points when performing significance tests***
- ***Validating models by using random subsets (bootstrapping, cross-validation)***

64. What is selection bias?

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample.

65. What are the types of biases that can occur during sampling?

1. ***Selection bias***
2. ***Undercoverage bias***
3. ***Survivorship bias***

66. What is survivorship bias?

Survivorship bias is the logical error of focusing on aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.

67. How do you work towards a random forest?

The underlying principle of this technique is that several weak learners combine to provide a strong learner. The steps involved are:

1. *Build several decision trees on bootstrapped training samples of data*
2. *On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates out of all pp predictors*
3. *Rule of thumb: At each split m=p√m=p*
4. *Predictions: At the majority rule*

This exhaustive list is sure to strengthen your preparation for data science interview questions.

68. What is a bias-variance trade-off?

Bias: Due to an oversimplification of a Machine Learning Algorithm, an error occurs in our model, which is known as Bias. This can lead to an issue of underfitting and might lead to oversimplified assumptions at the model training time to make target functions easier and simpler to understand.

Some of the popular machine learning algorithms which are low on the bias scale are -

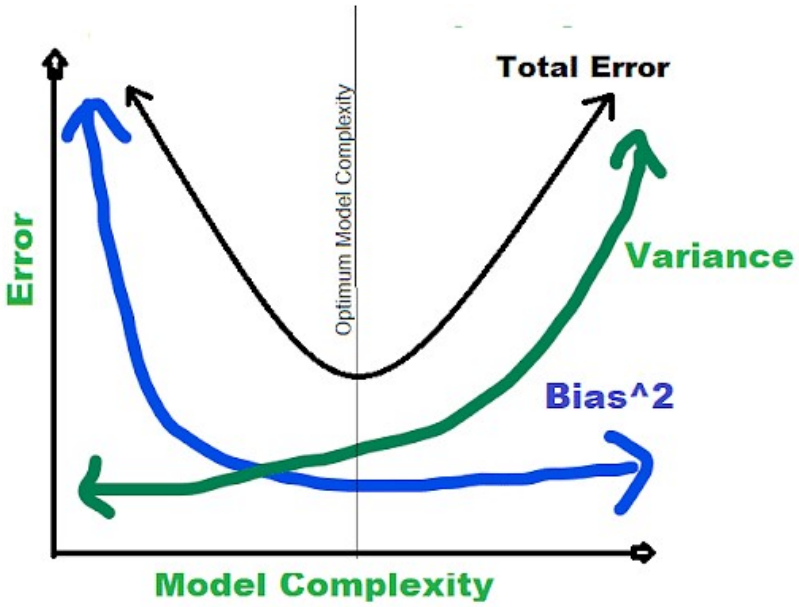
Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees.

Algorithms that are high on the bias scale -

Logistic Regression and Linear Regression.

Variance: Because of a complex machine learning algorithm, a model performs really badly on a test data set as the model learns even noise from the training data set. This error that occurs in the Machine Learning model is called Variance and can generate overfitting and hyper-sensitivity in Machine Learning models.

While trying to get over bias in our model, we try to increase the complexity of the machine learning algorithm. Though it helps in reducing the bias, after a certain point, it generates an overfitting effect on the model hence resulting in hyper-sensitivity and high variance.



Bias-Variance trade-off: To achieve the best performance, the main target of a supervised machine learning algorithm is to have low variance and bias.

The following things are observed regarding some of the popular machine learning algorithms -

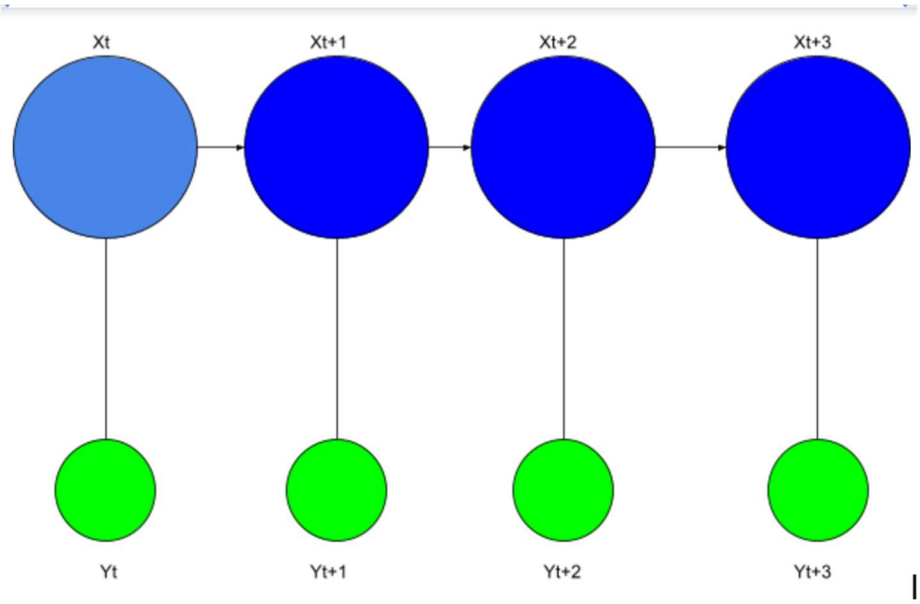
- *The Support Vector Machine algorithm (SVM) has high variance and low bias. In order to change the trade-off, we can increase the parameter C. The C parameter results in a decrease in the variance and an increase in bias by influencing the margin violations allowed in training datasets.*
- *In contrast to the SVM, the K-Nearest Neighbors (KNN) Machine Learning algorithm has a high variance and low bias. To change the trade-off of this algorithm, we can increase the prediction influencing neighbors by increasing the K value, thus increasing the model bias.*

69. Describe Markov chains?

Markov Chains defines that a state's future probability depends only on its current state.

Markov chains belong to the Stochastic process type category.

The below diagram explains a step-by-step model of the Markov Chains whose output depends on their current state.



A perfect example of the Markov Chains is the system of word recommendation. In this system, the model recognizes and recommends the next word based on the immediately previous word and not anything before that. The Markov Chains take the previous paragraphs that were similar to training data-sets and generates the recommendations for the current paragraphs accordingly based on the previous word.

70. Why is R used in Data Visualization?

R is widely used in Data Visualizations for the following reasons-

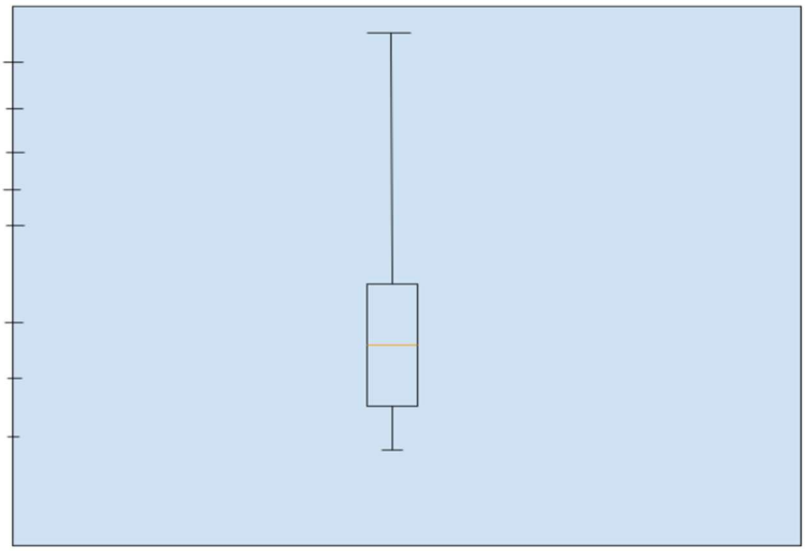
- *We can create almost any type of graph using R.*
- *R has multiple libraries like lattice, ggplot2, leaflet, etc., and so many inbuilt functions as well.*
- *It is easier to customize graphics in R compared to Python.*
- *R is used in feature engineering and in exploratory data analysis as well.*

71. What is the difference between a box plot and a histogram?

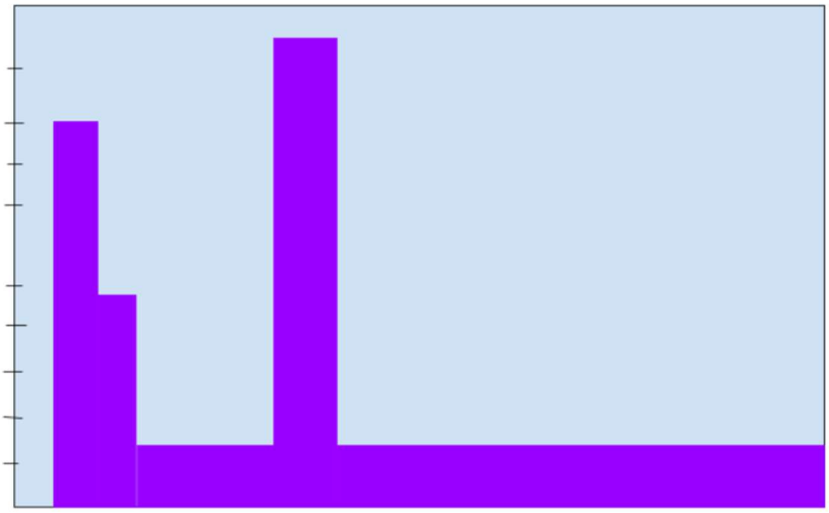
The frequency of a certain feature's values is denoted visually by both box plots

and histograms.

Boxplots are more often used in comparing several datasets and compared to histograms, take less space and contain fewer details. Histograms are used to know and understand the probability distribution underlying a dataset.



The diagram above denotes a boxplot of a dataset.



72. What does NLP stand for?

NLP is short for Natural Language Processing. It deals with the study of how computers learn a massive amount of textual data through programming. A few popular examples of NLP are Stemming, Sentimental Analysis, Tokenization, removal of stop words, etc.

73. Difference between an error and a residual error

The difference between a residual error and error are defined below -

Error	Residual Error

The difference between the actual value and the predicted value is called an error. *Some of the popular means of calculating data science errors are -* • *Root Mean Squared Error (RMSE)* • *Mean Absolute Error (MAE)* • *Mean Squared Error (MSE)*	*The difference between the arithmetic mean of a group of values and the observed group of values is called a residual error.*
An error is generally unobservable.	*A residual error can be represented using a graph.*
A residual error is used to show how the sample population data and the observed data differ from each other.	*An error is how actual population data and observed data differ from each other.*

74. Difference between Normalisation and Standardization

Standardization	*Normalization*
• *The technique of converting data in such a way that it is normally distributed and has a standard deviation of 1 and a mean of 0.*	• *The technique of converting all data values to lie between 1 and 0 is known as Normalization.*

	This is also known as min-max scaling.
• *Standardization takes care that the standard normal distribution is followed by the data.*	• *The data returning into the 0 to 1 range is taken care of by Normalization.*
• *Normalization formula -* *X' = (X - Xmin) / (Xmax - Xmin)* *Here,* *Xmin - feature's minimum value,* *Xmax - feature's maximum value.*	• *Standardization formula -* *X' = (X - μ) / σ*

75. Difference between Point Estimates and Confidence Interval

Confidence Interval: A range of values likely containing the population parameter is given by the confidence interval. Further, it even tells us how likely that particular interval can contain the population parameter. The Confidence Coefficient (or Confidence level) is denoted by 1-alpha, which gives the probability or likeness. The level of significance is given by alpha.

Point Estimates: An estimate of the population parameter is given by a particular value called the point estimate. Some popular methods used to derive Population Parameters' Point estimators are - Maximum Likelihood estimator and the Method of Moments.

To conclude, the bias and variance are inversely proportional to each other, i.e., an increase in bias results in a decrease in the variance, and an increase in variance results in a decrease in bias.

One-on-One Data Science Interview Questions

To crack a data science interview is no walk in the park. It requires in-depth knowledge and expertise in various topics. Furthermore, the projects that you have worked on can significantly boost your potential in a lot of interviews. In order to help you with your interviews, we have compiled a set of questions for you to relate to. Since data science is an extensive field, there are no limitations on the type of questions that can

be inquired. With that being said, you can answer each of these questions depending on the projects you have worked on and the industries you have been in. Try to answer each one of these sample questions and then share your answer with us through the comments.

Pro Tip: No matter how basic a question may seem, always try to view it from a technical perspective and use each question to demonstrate your unique technical skills and abilities.

76. Which is your favorite machine learning algorithm and why?

77. Which according to you is the most important skill that makes a good data scientist?

78. Why do you think data science is so popular today?

79. Explain the most challenging data science project that you worked on.

80. How do you usually prefer working on a project - individually, small team, or large team?

81. Based on your experience in the industry, tell me about your top 5 predictions for the next 10 years.

82. What are some unique skills that you can bring to the team as a data scientist?

83. Were you always in the data science field? If not, what made you change your career path and how did you upgrade your skills?

84. If we give you a random data set, how will you figure out whether it suits the business needs or not?

85. Given a chance, if you could pick a career other than being a data scientist, what would you choose?

86. Given the constant change in the data science field, how quickly can you adapt to new technologies?

87. Have you ever been in a conflict with your colleagues regarding different strategies to go about a project? How were you able to resolve it?

88. Can you break down an algorithm you have used on a recent project?

89. What tools did you use in your last project and why?

90. Think of the last technical problem that you solved. If you had no limitations with the project's budget, what would be the first thing you would do to solve the same problem?

91. When you are assigned multiple projects at the same time, how best do you organize your time?

92. Tell me about a time when your project didn't go according to plan and what you learned from it.

93. Have you ever created an original algorithm? How did you go about doing that and for what purpose?

94. What is your most favored strategy to clean a big data set and why?

95. Do you contribute to any open source projects?