DATA SCIENCE INTERVIEW QUESTIONS

What is Data Science?

Data Science is essentially a combination of algorithms, tools, methodologies, and machine learning techniques that help in finding hidden patterns from the given raw data.

What is logistic regression in Data Science?

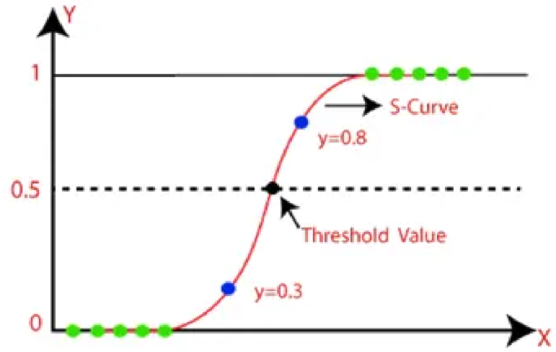
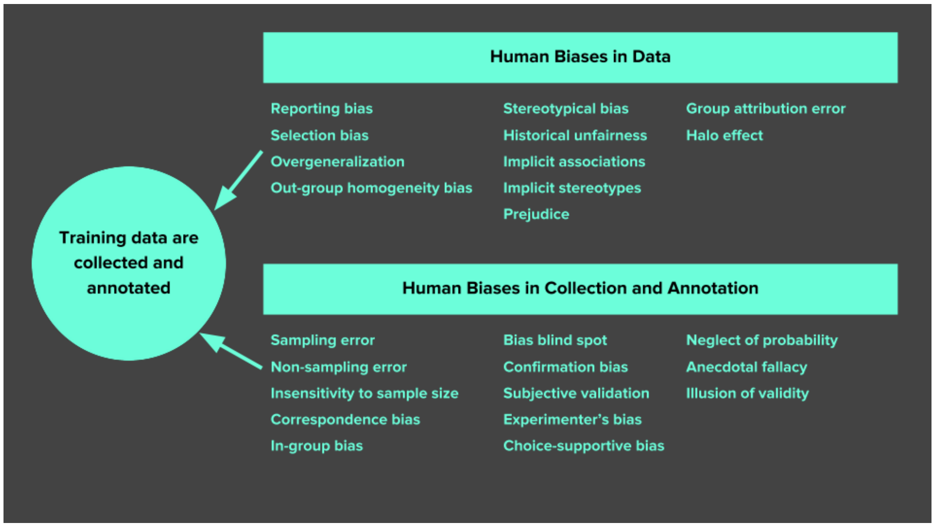


Fig 1:- Logistic Regression

Source: Javapoint

Logistic Regression is also called the logit model. It is a method to forecast the binary outcome from a linear combination of predictor variables.

Name three types of biases that can occur during sampling



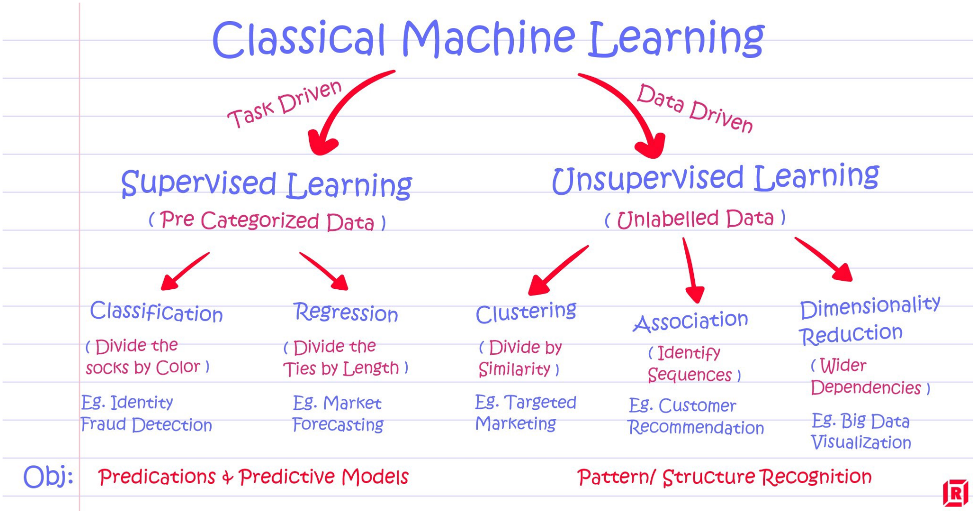
In the sampling process, there are three types of biases, which are:

- Selection bias
- Under coverage bias
- Survivorship bias

Discuss Decision Tree algorithm

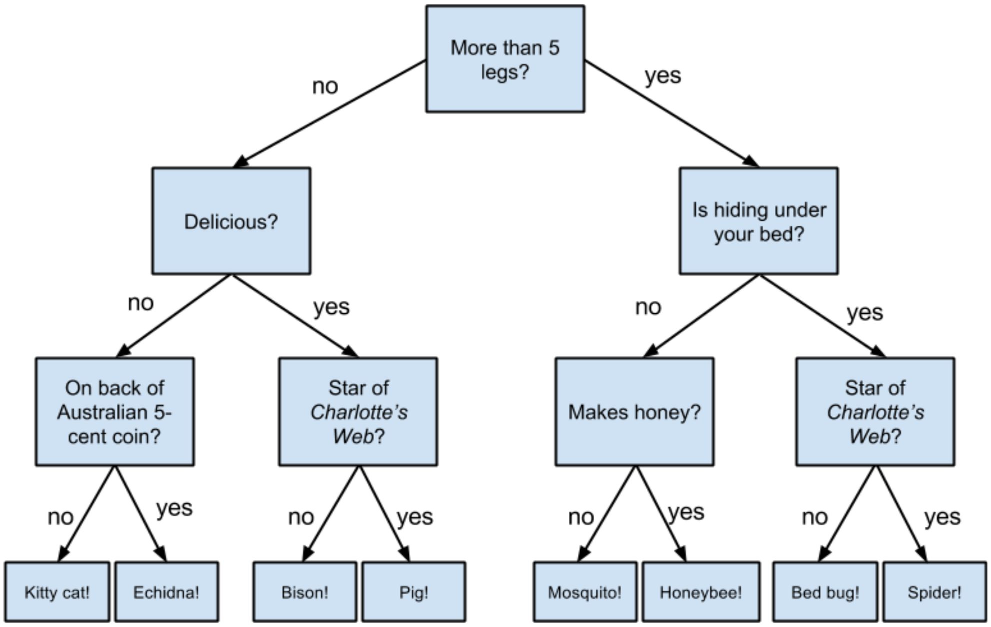
A decision tree is a popular supervised machine learning algorithm. It is mainly used for Regression and Classification. It breaks down a dataset into smaller subsets. The decision tree is able to handle both categorical and numerical data.

What are the differences between supervised and unsupervised learning?

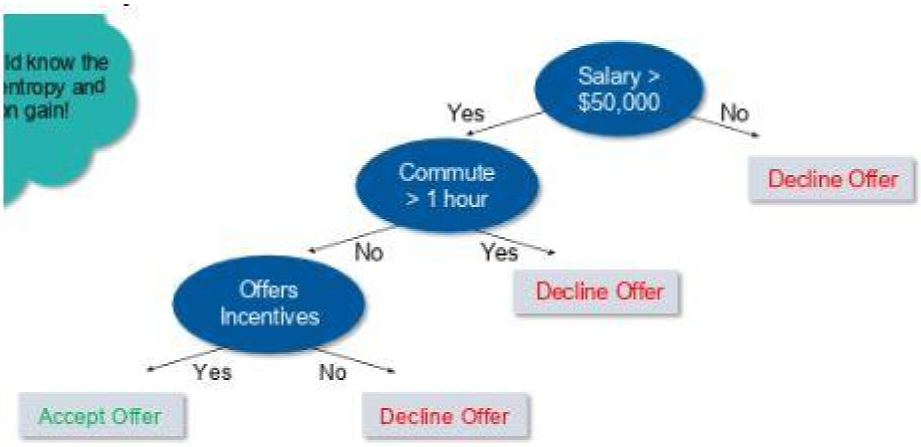


Supervised Learning	Unsupervised Learning
● Uses known and labeled data as input ● Supervised learning has a feedback mechanism ● The most commonly used supervised learning algorithms are decision trees, logistic regression, and support vector machine	● Uses unlabeled data as input ● Unsupervised learning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

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



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

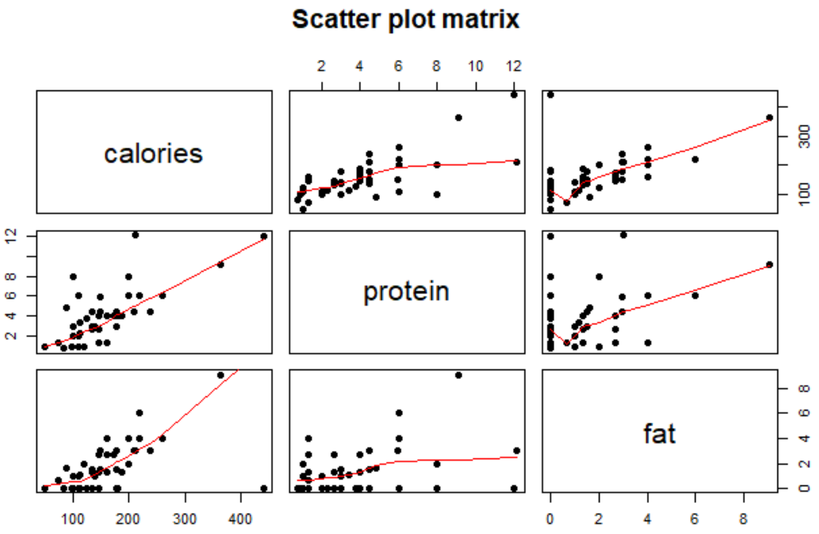
- Salary is greater than $50,000
- The commute is less than an hour
- Incentives are offered

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

Differentiate between univariate, bivariate, and multivariate analysis.



Univariate

Example: height of students

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.

Example: temperature and ice cream sales in the summer season

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

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

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

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

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

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).

What are dimensionality reduction and its benefits?

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.

How should you maintain a deployed model?

Monitor

Evaluate

Compare

Rebuild

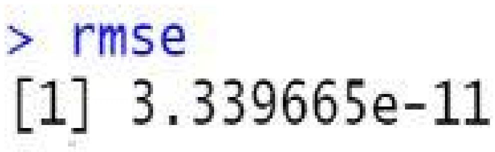
What are recommender systems?

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

Content-based Filtering

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



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

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

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.

How can outlier values be treated?

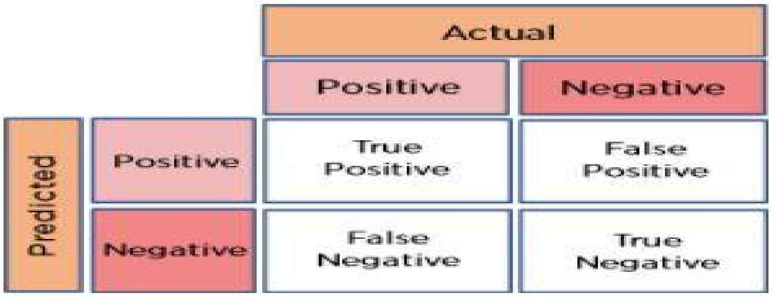
You can drop outliers only if it is a garbage value.

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.

How can you calculate accuracy using a confusion matrix?

Consider this confusion matrix:



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

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

= (262 + 347) / 650

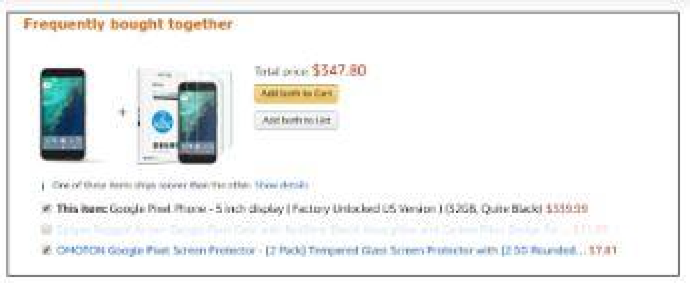
= 609 / 650

= 0.93

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

'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.



Frequently bought together

Total price: $347.80

One of these items ships sooner than the other. Show details

☑ **This item:** Google Pixel Phone - 5 inch display | Factory Unlocked US Version | (32GB, Quite Black) $339.99

☐ [illegible]

☑ OMOTON Google Pixel Screen Protector - [2 Pack] Tempered Glass Screen Protector with [2.5D Rounded... $7.81

Explain Recommender Systems?

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

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

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

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

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

What is Power Analysis?

The power analysis is an integral part of the experimental design. It helps you to determine the sample size required to find out the effect of a given size from a cause with a specific level of assurance.

Explain Collaborative filtering

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

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.

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.

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.

State the difference between the expected value and mean value

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.

What is Ensemble Learning?

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 depending upon the last classification. Boosting decreases the bias error and helps you to build strong predictive models.

Explain Eigenvalue and Eigenvector

Eigenvectors are for understanding linear transformations. Data scientists need to calculate the eigenvectors for a covariance matrix or correlation.

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.

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.

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 depending upon the error rate obtained in the previous epoch.

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.

Explain the difference between Data Science and Data Analytics

The main difference between the two is that the data scientists have more technical knowledge than business analysts. Moreover, they don't need an understanding of the business required for data visualization.

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.

When underfitting occurs in a static model?

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

What is reinforcement learning?

Reinforcement Learning is a learning mechanism about how to map situations to actions.

What is precision?

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

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.

State the difference between a Validation Set and a Test Set

A Validation set is 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.

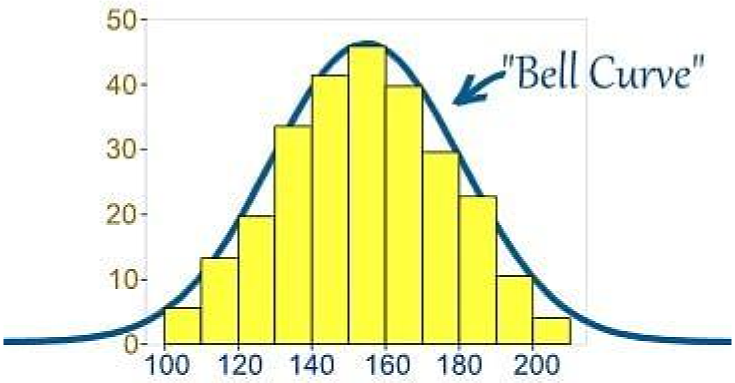
While a Test Set is used for testing or evaluating the performance of a trained machine learning model.

Why does data cleaning play a vital role in the analysis?

Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because - as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources.

What do you understand by the term Normal Distribution?

The random variables are distributed in the form of a symmetrical bell shaped curve.



What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X.

What are Interpolation and Extrapolation?

Estimating a value from 2 known values from a list of values is Interpolation. Extrapolation is approximating a value by extending a known set of values or facts.

What is power analysis?

An experimental design technique for determining the effect of a given sample size.

What is K-means? How can you select K for K-means?

What is Collaborative filtering?

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources, and multiple agents.

What is the difference between Cluster and Systematic Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. A cluster sample is a probability sample where each sampling unit is a collection or cluster of elements.

What is the benefit of shuffling a training dataset when using a batch gradient descent algorithm for optimizing a neural network?

Mini batch gradient descent is a compromise between Batch gradient descent and Stochastic gradient descent where small batches of data are considered to take each step of gradient descent. In both Batch gradient descent and Mini Batch gradient descent, shuffling of data after each epoch is crucial.

During analysis, how do you treat missing values?

- Understand the problem statement, understand the data and then give the answer.Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.
- If it is a categorical variable, the default value is assigned. The missing value is assigned a default value.
- If you have a distribution of data coming, for normal distribution give the mean value.