**Communication (5 questions)**

1. Explain to me a technical concept related to the role that you’re interviewing for.

2. Introduce me to something you’re passionate about.

3. How would you explain an A/B test to an engineer with no statistics background? A linear regression?

A/B testing, or more broadly, multivariate testing, is the testing of different elements of a user's experience to determine which variation helps the business achieve its goal more effectively (i.e. increasing conversions, etc.)  This can be copy on a web site, button colors, different user interfaces, different email subject lines, calls to action, offers, etc.

4. How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean? [link](https://www.quora.com/What-is-a-confidence-interval-in-laymans-terms)

5. How would you explain to a group of senior executives why data is important?

**Data Analysis (27 questions)**

1. (Given a Dataset) Analyze this dataset and tell me what you can learn from it.

2. What is R2? What are some other metrics that could be better than R2 and why?

goodness of fit measure. variance explained by the regression / total variance

the more predictors you add the higher R^2 becomes.

hence use adjusted R^2 which adjusts for the degrees of freedom

or train error metrics

3. What is the curse of dimensionality?

High dimensionality makes clustering hard, because having lots of dimensions means that everything is "far away" from each other.For example, to cover a fraction of the volume of the data we need to capture a very wide range for each variable as the number of variables increases All samples are close to the edge of the sample. And this is a bad news because prediction is much more difficult near the edges of the training sample. The sampling density decreases exponentially as p increases and hence the data becomes much more sparse without significantly more data. We should conduct PCA to reduce dimensionality

4. Is more data always better?

Statistically,It depends on the quality of your data, for example, if your data is biased, just getting more data won’t help.It depends on your model. If your model suffers from high bias, getting more data won’t improve your test results beyond a point. You’d need to add more features, etc.

Practically,Also there’s a tradeoff between having more data and the additional storage, computational power, memory it requires. Hence, always think about the cost of having more data.

5. What are advantages of plotting your data before per- forming analysis?

Data sets have errors.  You won't find them all but you might find some. That 212 year old man. That 9 foot tall woman.Variables can have skewness, outliers etc.  Then the arithmetic mean might not be useful. Which means the standard deviation isn't useful.Variables can be multimodal!  If a variable is multimodal then anything based on its mean or median is going to be suspect.

6. How can you make sure that you don’t analyze something that ends up meaningless?

Proper exploratory data analysis.

In every data analysis task, there's the exploratory phase where you're just graphing things, testing things on small sets of the data, summarizing simple statistics, and getting rough ideas of what hypotheses you might want to pursue further.Then there's the exploitatory phase, where you look deeply into a set of hypotheses. The exploratory phase will generate lots of possible hypotheses, and the exploitatory phase will let you really understand a few of them. Balance the two and you'll prevent yourself from wasting time on many things that end up meaningless, although not all.

7. What is the role of trial and error in data analysis? What is the the role of making a hypothesis before diving in?

data analysis is a repetition of setting up a new hypothesis and trying to refute the null hypothesis.

The scientific method is eminently inductive: we elaborate a hypothesis, test it and refute it or not. As a result, we come up with new hypotheses which are in turn tested and so on. This is an iterative process, as science always is.

8. How can you determine which features are the most im- portant in your model?

run the features though a Gradient Boosting Machine or Random Forest to generate plots of relative importance and information gain for each feature in the ensembles.Look at the variables added in forward variable selection

9. How do you deal with some of your predictors being missing?

Remove rows with missing values - This works well if 1) the values are missing randomly (see [Vinay Prabhu's answer](https://www.quora.com/How-can-I-deal-with-missing-values-in-a-predictive-model/answer/Vinay-Prabhu-7) for more details on this) 2) if you don't lose too much of the dataset after doing so.

Build another predictive model to predict the missing values - This could be a whole project in itself, so simple techniques are usually used here.

Use a model that can incorporate missing data - Like a random forest, or any tree-based method.

10. You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?

Multicollinearity refers to a situation in which two or more explanatory variables in a [multiple regression](https://en.wikipedia.org/wiki/Multiple_regression) model are highly linearly related. Leave the model as is, despite multicollinearity. The presence of multicollinearity doesn't affect the efficiency of extrapolating the fitted model to new data provided that the predictor variables follow the same pattern of multicollinearity in the new data as in the data on which the regression model is based.

principal component regression

11. Let’s say you’re given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?

PCA

12. Now you have a feasible amount of predictors, but you’re fairly sure that you don’t need all of them. How would you perform feature selection on the dataset?

ridge / lasso / elastic net regression Univariate Feature Selection where a statistical test is applied to each feature individually. You retain only the best features according to the test outcome scores

"Recursive Feature Elimination": First, train a model with all the feature and evaluate its performance on held out data. Then drop let say the 10% weakest features (e.g. the feature with least absolute coefficients in a linear model) and retrain on the remaining features. Iterate until you observe a sharp drop in the predictive accuracy of the model.

13. Your linear regression didn’t run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?

p > n.

If some of the explanatory variables are perfectly correlated (positively or negatively) then the coefficients would not be unique.

14. You run your regression on different subsets of your data, and find that in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

The dataset might be heterogeneous. In which case, it is recommended to cluster datasets into different subsets wisely, and then draw different models for different subsets. Or, use models like non parametric models (trees) which can deal with heterogeneity quite nicely.

15. What is the main idea behind ensemble learning? If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse?

The assumption is that a group of weak learners can be combined to form a strong learner.

Hence the combined model is expected to perform better than an individual model.

Assumptions: average out biases reduce variance Bagging works because some underlying learning algorithms are unstable: slightly different inputs leads to very different outputs. If you can take advantage of this instability by running multiple instances, it can be shown that the reduced instability leads to lower error. If you want to understand why, the original bagging paper( [http://www.springerlink.com/cont...](http://www.springerlink.com/content/l4780124w2874025/)) has a section called "why bagging works"

Boosting works because of the focus on better defining the "decision edge". By reweighting examples near the margin (the positive and negative examples) you get a reduced error (see <http://citeseerx.ist.psu.edu/vie>...)

Use the outputs of your models as inputs to a meta-model.

For example, if you're doing binary classification, you can use all the probability outputs of your individual models as inputs to a final logistic regression (or any model, really) that can combine the probability estimates.

One very important point is to make sure that the output of your models are out-of-sample predictions. This means that the predicted value for any row in your dataframe should NOT depend on the actual value for that row.

16. Given that you have wi  data in your o ce, how would you determine which rooms and areas are underutilized and overutilized?

If the data is more used in one room, then that one is over utilized! Maybe account for the room capacity and normalize the data.

17. How could you use GPS data from a car to determine the quality of a driver?

18. Given accelerometer, altitude, and fuel usage data from a car, how would you determine the optimum acceleration pattern to drive over hills?

19. Given position data of NBA players in a season’s games, how would you evaluate a basketball player’s defensive ability?

20. How would you quantify the influence of a Twitter user?

like page rank with each user corresponding to the webpages and linking to the page equivalent to following.

21. Given location data of golf balls in games, how would construct a model that can advise golfers where to aim?

22. You have 100 mathletes and 100 math problems. Each mathlete gets to choose 10 problems to solve. Given data on who got what problem correct, how would you rank the problems in terms of di culty?

One way you could do this is by storing a "skill level" for each user and a "difficulty level" for each problem.  We assume that the probability that a user solves a problem only depends on the skill of the user and the difficulty of the problem.\*  Then we maximize the likelihood of the data to find the hidden skill and difficulty levels.

The Rasch model for dichotomous data takes the form:  
{\displaystyle \Pr\{X\_{ni}=1\}={\frac {\exp({\beta \_{n}}-{\delta \_{i}})}{1+\exp({\beta \_{n}}-{\delta \_{i}})}},}  
where  is the ability of person  and  is the difficulty of item}.

23. You have 5000 people that rank 10 sushis in terms of saltiness. How would you aggregate this data to estimate the true saltiness rank in each sushi?

Some people would take the mean rank of each sushi.  If I wanted something simple, I would use the median, since ranks are (strictly speaking) ordinal and not interval, so adding them is a bit risque (but people do it all the time and you probably won't be far wrong).

24. Given data on congressional bills and which congressional representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most liberal? Most republican? Most bipartisan?

collaborative filtering. you have your votes and we can calculate the similarity for each representatives and select the most similar representative

for liberal and republican parties, find the mean vector and find the representative closest to the center point

25. How would you come up with an algorithm to detect plagiarism in online content?

reduce the text to a more compact form (e.g. fingerprinting, bag of words) then compare those with other texts by calculating the similarity

26. You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters to include?

KNN

choose a small value of k that still has a low SSE (elbow method)

<https://bl.ocks.org/rpgove/0060ff3b656618e9136b>

27. Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach this problem?

[collaborative filtering](https://en.wikipedia.org/wiki/Collaborative_filtering)

**Predictive Modeling (19 questions)**

1. (Given a Dataset) Analyze this dataset and give me a model that can predict this response variable.

Start by fitting a simple model (multivariate regression, logistic regression), do some feature engineering accordingly, and then try some complicated models. Always split the dataset into train, validation, test dataset and use cross validation to check their performance.

Determine if the problem is classification or regression

Favor simple models that run quickly and you can easily explain.

Mention cross validation as a means to evaluate the model.

Plot and visualize the data.

2. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

The model that has high training accuracy might have low test accuracy. Without further knowledge, it is hard to know which dataset represents the population data and thus the generalizability of the algorithm is hard to measure. This should be mitigated by repeated splitting of train vs test dataset (as in cross validation).

When there is a change in data distribution, this is called the dataset shift. If the train and test data has a different distribution, then the classifier would likely overfit to the train data.

This issue can be overcome by using a more general learning method.

This can occur when:

P(y|x) are the same but P(x) are different. (covariate shift)

P(y|x) are different. (concept shift)

The causes can be:

Training samples are obtained in a biased way. (sample selection bias)

Train is different from test because of temporal, spatial changes. (non-stationary environments)

Solution to covariate shift

importance weighted cv

3. What are some ways I can make my model more robust to outliers?

We can have regularization such as L1 or L2 to reduce variance (increase bias).

Changes to the algorithm:

Use tree-based methods instead of regression methods as they are more resistant to outliers. For statistical tests, use non parametric tests instead of parametric ones.

Use robust error metrics such as MAE or Huber Loss instead of MSE.

Changes to the data:

Winsorizing the data

Transforming the data (e.g. log)

Remove them only if you’re certain they’re anomalies not worth predicting

4. What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?

MSE is more strict to having outliers. MAE is more robust in that sense, but is harder to fit the model for because it cannot be numerically optimized. So when there are less variability in the model and the model is computationally easy to fit, we should use MAE, and if that’s not the case, we should use MSE.

MSE: easier to compute the gradient, MAE: linear programming needed to compute the gradient

MAE more robust to outliers. If the consequences of large errors are great, use MSE

MSE corresponds to maximizing likelihood of Gaussian random variables

5. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

Accuracy: proportion of instances you predict correctly. Pros: intuitive, easy to explain, Cons: works poorly when the class labels are imbalanced and the signal from the data is weak

AUROC: plot fpr on the x axis and tpr on the y axis for different threshold. Given a random positive instance and a random negative instance, the AUC is the probability that you can identify who's who. Pros: Works well when testing the ability of distinguishing the two classes, Cons: can’t interpret predictions as probabilities (because AUC is determined by rankings), so can’t explain the uncertainty of the model

logloss/deviance: Pros: error metric based on probabilities, Cons: very sensitive to false positives, negatives

When there are more than 2 groups, we can have k binary classifications and add them up for logloss. Some metrics like AUC is only applicable in the binary case.

6. What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What’s the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)

Things to look at: N, P, linearly seperable?, features independent?, likely to overfit?, speed, performance, memory usage

Logistic Regression

features roughly linear, problem roughly linearly separable

robust to noise, use l1,l2 regularization for model selection, avoid overfitting

the output come as probabilities

efficient and the computation can be distributed

can be used as a baseline for other algorithms

(-) can hardly handle categorical features

SVM

with a nonlinear kernel, can deal with problems that are not linearly separable

(-) slow to train, for most industry scale applications, not really efficient

Naive Bayes

computationally efficient when P is large by alleviating the curse of dimensionality

works surprisingly well for some cases even if the condition doesn’t hold

with word frequencies as features, the independence assumption can be seen reasonable. So the algorithm can be used in text categorization

(-) conditional independence of every other feature should be met

Tree Ensembles

good for large N and large P, can deal with categorical features very well

non parametric, so no need to worry about outliers

GBT’s work better but the parameters are harder to tune

RF works out of the box, but usually performs worse than GBT

Deep Learning

works well for some classification tasks (e.g. image)

used to squeeze something out of the problem

7. What is regularization and where might it be helpful? What is an example of using regularization in a model?

Regularization is useful for reducing variance in the model, meaning avoiding overfitting . For example, we can use L1 regularization in Lasso regression to penalize large coefficients.

8. Why might it be preferable to include fewer predictors over many?

When we add irrelevant features, it increases model's tendency to overfit because those features introduce more noise. When two variables are correlated, they might be harder to interpret in case of regression, etc.

curse of dimensionality

adding random noise makes the model more complicated but useless

computational cost

Ask someone for more details.

9. Given training data on tweets and their retweets, how would you predict the number of retweets of a given tweet after 7 days after only observing 2 days worth of data?

Build a time series model with the training data with a seven day cycle and then use that for a new data with only 2 days data.

Ask someone for more details.

Build a regression function to estimate the number of retweets as a function of time t

to determine if one regression function can be built, see if there are clusters in terms of the trends in the number of retweets

if not, we have to add features to the regression function

features + # of retweets on the first and the second day -> predict the seventh day

<https://en.wikipedia.org/wiki/Dynamic_time_warping>

10. How could you collect and analyze data to use social media to predict the weather?

We can collect social media data using twitter, Facebook, instagram API’s. Then, for example, for twitter, we can construct features from each tweet, e.g. the tweeted date, number of favorites, retweets, and of course, the features created from the tweeted content itself. Then use a multi variate time series model to predict the weather.

Ask someone for more details.

11. How would you construct a feed to show relevant content for a site that involves user interactions with items?

We can do so using building a recommendation engine. The easiest we can do is to show contents that are popular other users, which is still a valid strategy if for example the contents are news articles. To be more accurate, we can build a content based filtering or collaborative filtering. If there’s enough user usage data, we can try collaborative filtering and recommend contents other similar users have consumed. If there isn’t, we can recommend similar items based on vectorization of items (content based filtering).

12. How would you design the people you may know feature on LinkedIn or Facebook?

Find strong unconnected people in weighted connection graph

Define similarity as how strong the two people are connected

Given a certain feature, we can calculate the similarity based on

friend connections (neighbors)

Check-in’s people being at the same location all the time.

same college, workplace

Have randomly dropped graphs test the performance of the algorithm

ref. News Feed Optimization

Affinity score: how close the content creator and the users are

Weight: weight for the edge type (comment, like, tag, etc.). Emphasis on features the company wants to promote

Time decay: the older the less important

13. How would you predict who someone may want to send a Snapchat or Gmail to?

for each user, assign a score of how likely someone would send an email to

the rest is feature engineering:

number of past emails, how many responses, the last time they exchanged an email, whether the last email ends with a question mark, features about the other users, etc.

Ask someone for more details.

People who someone sent emails the most in the past, conditioning on time decay.

14. How would you suggest to a franchise where to open a new store?

build a master dataset with local demographic information available for each location.

local income levels, proximity to traffic, weather, population density, proximity to other businesses

a reference dataset on local, regional, and national macroeconomic conditions (e.g. unemployment, inflation, prime interest rate, etc.)

any data on the local franchise owner-operators, to the degree the manager

identify a set of KPIs acceptable to the management that had requested the analysis concerning the most desirable factors surrounding a franchise

quarterly operating profit, ROI, EVA, pay-down rate, etc.

run econometric models to understand the relative significance of each variable

run machine learning algorithms to predict the performance of each location candidate

15. In a search engine, given partial data on what the user has typed, how would you predict the user’s eventual search query?

Based on the past frequencies of words shown up given a sequence of words, we can construct conditional probabilities of the set of next sequences of words that can show up (n-gram). The sequences with highest conditional probabilities can show up as top candidates.

To further improve this algorithm,

we can put more weight on past sequences which showed up more recently and near your location to account for trends

show your recent searches given partial data

16. Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

Based on frequency and amount of donations, graduation year, major, etc, construct a supervised regression (or binary classification) algorithm.

17. You’re Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?

Based on the past pickup location of passengers around the same time of the day, day of the week (month, year), construct

Ask someone for more details.

Based on the number of past pickups

account for periodicity (seasonal, monthly, weekly, daily, hourly)

special events (concerts, festivals, etc.) from tweets

18. How would you build a model to predict a March Madness bracket?

One vector each for team A and B. Take the difference of the two vectors and use that as an input to predict the probability that team A would win by training the model. Train the models using past tournament data and make a prediction for the new tournament by running the trained model for each round of the tournament

Some extensions:

Experiment with different ways of consolidating the 2 team vectors into one (e.g concantenating, averaging, etc)

Consider using a RNN type model that looks at time series data.

19. You want to run a regression to predict the probability of a flight delay, but there are flights with delays of up to 12 hours that are really messing up your model. How can you address this?

This is equivalent to making the model more robust to outliers.

See Q3.

**Probability (19 questions)**

1. Bobo the amoeba has a 25%, 25%, and 50% chance of producing 0, 1, or 2 o spring, respectively. Each of Bobo’s descendants also have the same probabilities. What is the probability that Bobo’s lineage dies out?

p=1/4+1/4p+1/2p^2 => p=1/2

2. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the proba- bility that you see at least one shooting star in the period of an hour?

1-(0.8)^4. Or, we can use Poisson processes

3. How can you generate a random number between 1 - 7 with only a die?

Launch it 3 times: each throw sets the nth bit of the result.

For each launch, if the value is 1-3, record a 0, else 1. The result is between 0 (000) and 7 (111), evenly spread (3 independent throw). Repeat the throws if 0 was obtained: the process stops on evenly spread values.

4. How can you get a fair coin toss if someone hands you a coin that is weighted to come up heads more often than tails?

Flip twice and if HT then H, TH then T.

5. You have an 50-50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?

more than two standard deviations

6. Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?

plug in the value to the CDF of the same random variable

7. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

1/3

8. You have a group of couples that decide to have children until they have their first girl, after which they stop having children. What is the expected gender ratio of the children that are born? What is the expected number of children each couple will have?

gender ratio is 1:1. Expected number of children is 2. let X be the number of children until getting a female (happens with prob 1/2). this follows a geometric distribution with probability 1/2

9. How many ways can you split 12 people into 3 teams of 4?

the outcome follows a multinomial distribution with n=12 and k=3. but the classes are indistinguishable

10. Your hash function assigns each object to a number between 1:10, each with equal probability. With 10 objects, what is the probability of a hash collision? What is the expected number of hash collisions? What is the expected number of hashes that are unused.

the probability of a hash collision: 1-(10!/10^10)

the expected number of hash collisions: 1-10\*(9/10)^10

the expected number of hashes that are unused: 10\*(9/10)^10

11. You call 2 UberX’s and 3 Lyfts. If the time that each takes to reach you is IID, what is the probability that all the Lyfts arrive first? What is the probability that all the UberX’s arrive first?

Lyfts arrive first: 2!\*3!/5!

Ubers arrive first: same

12. I write a program should print out all the numbers from 1 to 300, but prints out Fizz instead if the number is divisible by 3, Buzz instead if the number is divisible by 5, and FizzBuzz if the number is divisible by 3 and 5. What is the total number of numbers that is either Fizzed, Buzzed, or FizzBuzzed?

100+60-20=140

13. On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Alice and Bob randomly pick adjectives, what is the probability that they form a match?

24C5\*(1+5(24-5))/24C5\*24C5 = 4/1771

14. A lazy high school senior types up application and envelopes to n different colleges, but puts the applications randomly into the envelopes. What is the expected number of applications that went to the right college?

1

15. Let’s say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?

Shorter. Regression to the mean

16. What’s the expected number of coin flips until you get two heads in a row? What’s the expected number of coin flips until you get two tails in a row?

17. Let’s say we play a game where I keep flipping a coin until I get heads. If the first time I get heads is on the nth coin, then I pay you 2n-1 dollars. How much would you pay me to play this game?

less than $3

18. You have two coins, one of which is fair and comes up heads with a probability 1/2, and the other which is biased and comes up heads with probability 3/4. You randomly pick coin and flip it twice, and get heads both times. What is the probability that you picked the fair coin?

4/13

19. You have a 0.1% chance of picking up a coin with both heads, and a 99.9% chance that you pick up a fair coin. You flip your coin and it comes up heads 10 times. What’s the chance that you picked up the fair coin, given the information that you observed?

Events: F = "picked a fair coin", T = "10 heads in a row"

(1) P(F|T) = P(T|F)P(F)/P(T) (Bayes formula)

(2) P(T) = P(T|F)P(F) + P(T|¬F)P(¬F) (total probabilities formula)

Injecting (2) in (1): P(F|T) = P(T|F)P(F)/(P(T|F)P(F) + P(T|¬F)P(¬F)) = 1 / (1 + P(T|¬F)P(¬F)/(P(T|F)P(F)))

Numerically: 1/(1 + 0.001 \* 2^10 /0.999).

With 2^10 ≈ 1000 and 0.999 ≈ 1 this simplifies to 1/2

20. What is a P-Value ?

The probability to obtain a similar or more extreme result than observed when the null hypothesis is assumed.

⇒ If the p-value is small, the null hypothesis is unlikely

Programming (14 questions)

1. Write a function to calculate all possible assignment vectors of 2n users, where n users are assigned to group 0 (control), and n users are assigned to group 1 (treatment).

Recursive programming (sol in code)

2. Given a list of tweets, determine the top 10 most used hashtags.

Store all the hashtags in a dictionary and get the top 10 values

3. Program an algorithm to find the best approximate solution to the knapsack problem1 in a given time.

Greedy solution (add the best v/w as much as possible and move on to the next)

4. Program an algorithm to find the best approximate solution to the travelling salesman problem2 in a given time.

Greedy

5. You have a stream of data coming in of size n, but you don’t know what n is ahead of time. Write an algorithm that will take a random sample of k elements. Can you write one that takes O(k) space?

<https://en.wikipedia.org/wiki/Reservoir_sampling>

6. Write an algorithm that can calculate the square root of a number.

<https://www.quora.com/What-is-the-method-to-calculate-a-square-root-by-hand?redirected_qid=664405>

<https://en.wikipedia.org/wiki/Newton's_method#Square_root_of_a_number>

7. Given a list of numbers, can you return the outliers?

sort then select the highest and the lowest 2.5%

8. When can parallelism make your algorithms run faster?

When could it make your algorithms run slower?

Ask someone for more details.

compute in parallel when communication cost < computation cost

ensemble trees

minibatch

cross validation

forward propagation

minibatch

not suitable for online learning

9. What are the different types of joins? What are the differences between them?

(INNER) JOIN: Returns records that have matching values in both tables LEFT (OUTER) JOIN: Return all records from the left table, and the matched records from the right table RIGHT (OUTER) JOIN: Return all records from the right table, and the matched records from the left table FULL (OUTER) JOIN: Return all records when there is a match in either left or right table

10. Why might a join on a subquery be slow? How might you speed it up?

Change the subquery to a join.

11. Describe the difference between primary keys and foreign keys in a SQL database.

Primary keys are columns whose value combinations must be unique in a specific table so that each row can be referenced uniquely. Foreign keys are columns that references columns (often primary keys) in other tables.

12. Given a COURSES table with columns course\_id and course\_name, a FACULTY table with columns faculty\_id and faculty\_name, and a COURSE\_FACULTY table with columns faculty\_id and course\_id, how would you return a list of faculty who teach a course given the name of a course?

select faculty\_name from faculty\_id c join (select faculty\_id from (select course\_id from COURSES where course\_name=xxx) as a join COURSE\_FACULTY b on a.course\_id = b.course\_id) d on c.faculty\_id = d.faculty\_id

13. Given a IMPRESSIONS table with ad\_id, click (an indicator that the ad was clicked), and date, write a SQL query that will tell me the click-through-rate of each ad by month.

select ad\_id, average(click) from IMPRESSIONS group by ad\_id, month(date)

14. Write a query that returns the name of each department and a count of the number of employees in each:

EMPLOYEES containing: Emp\_ID (Primary key) and Emp\_Name  
EMPLOYEE\_DEPT containing: Emp\_ID (Foreign key) and Dept\_ID (Foreign key)  
DEPTS containing: Dept\_ID (Primary key) and Dept\_Name

select Dept\_Name, count(1) from DEPTS a right join EMPLOYEE\_DEPT b on a.Dept\_id = b.Dept\_id group by Dept\_Name

Statistical Inference (15 questions)

1. In an A/B test, how can you check if assignment to the various buckets was truly random?

Plot the distributions of multiple features for both A and B and make sure that they have the same shape. More rigorously, we can conduct a permutation test to see if the distributions are the same.

MANOVA to compare different means

2. What might be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?

Verify the sampling algorithm is random.

3. What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?

The user might not act the same suppose had they not seen the other bucket. You are essentially adding additional variables of whether the user peeked the other bucket, which are not random across groups.

4. What would be some issues if blogs decide to cover one of your experimental groups?

Same as the previous question. The above problem can happen in larger scale.

5. How would you conduct an A/B test on an opt-in feature?

Ask someone for more details.

6. How would you run an A/B test for many variants, say 20 or more?

one control, 20 treatment, if the sample size for each group is big enough.

Ways to attempt to correct for this include changing your confidence level (e.g. Bonferroni Correction) or doing family-wide tests before you dive in to the individual metrics (e.g. Fisher's Protected LSD).

7. How would you run an A/B test if the observations are extremely right-skewed?

lower the variability by modifying the KPI

cap values

percentile metrics

log transform

<https://www.quora.com/How-would-you-run-an-A-B-test-if-the-observations-are-extremely-right-skewed>

8. I have two different experiments that both change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?

exclusive -> ok

9. What is a p-value? What is the di erence between type-1 and type-2 error?

type-1 error: rejecting Ho when Ho is true

type-2 error: not rejecting Ho when Ha is true

10. You are AirBnB and you want to test the hypothesis that a greater number of photographs increases the chances that a buyer selects the listing. How would you test this hypothesis?

For randomly selected listings with more than 1 pictures, hide 1 random picture for group A, and show all for group B. Compare the booking rate for the two groups.

Ask someone for more details.

11. How would you design an experiment to determine the impact of latency on user engagement?

The best way I know to quantify the impact of performance is to isolate just that factor using a slowdown experiment, i.e., add a delay in an A/B test.

12. What is maximum likelihood estimation? Could there be any case where it doesn’t exist?

A method for parameter optimization (fitting a model). We choose parameters so as to maximize the likelihood function (how likely the outcome would happen given the current data and our model).

maximum likelihood estimation (MLE) is a method of [estimating](https://en.wikipedia.org/wiki/Estimator) the [parameters](https://en.wikipedia.org/wiki/Statistical_parameter) of a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) given observations, by finding the parameter values that maximize the [likelihood](https://en.wikipedia.org/wiki/Likelihood) of making the observations given the parameters. MLE can be seen as a special case of the [maximum a posteriori estimation](https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation) (MAP) that assumes a [uniform](https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)) [prior distribution](https://en.wikipedia.org/wiki/Prior_probability) of the parameters, or as a variant of the MAP that ignores the prior and which therefore is [unregularized](https://en.wikipedia.org/wiki/Regularization_(mathematics)).

for gaussian mixtures, non parametric models, it doesn’t exist

13. What’s the di erence between a MAP, MOM, MLE estima- tor? In which cases would you want to use each?

MAP estimates the posterior distribution given the prior distribution and data which maximizes the likelihood function. MLE is a special case of MAP where the prior is uninformative uniform distribution.

MOM sets moment values and solves for the parameters. MOM is not used much anymore because maximum likelihood estimators have higher probability of being close to the quantities to be estimated and are more often unbiased.

14. What is a confidence interval and how do you interpret it?

For example, 95% confidence interval is an interval that when constructed for a set of samples each sampled in the same way, the constructed intervals include the true mean 95% of the time.

if confidence intervals are constructed using a given confidence level in an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level.

15. What is unbiasedness as a property of an estimator? Is this always a desirable property when performing inference? What about in data analysis or predictive modeling?

Unbiasedness means that the expectation of the estimator is equal to the population value we are estimating. This is desirable in inference because the goal is to explain the dataset as accurately as possible. However, this is not always desirable for data analysis or predictive modeling as there is the bias variance tradeoff. We sometimes want to prioritize the generalizability and avoid overfitting by reducing variance and thus increasing bias.

**Statistics Interview Questions for Data Science**

1. Explain the central limit theorem.

2. What is the relevance of central limit theorem to a class of freshmen in the social sciences who hardly have any knowledge about statistics?

3. Given a dataset, show me how Euclidean Distance works in three dimensions.

4. How will you prevent overfitting when creating a statistical model?

**Frequently Asked Open Ended Machine Learning Interview Questions for Data Scientists**

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

2.In which libraries for Data Science in Python and R, does your strength lie?

3.What kind of data is important for specific business requirements and how, as a data scientist will you go about collecting that data?

4.Tell us about the biggest data set you have processed till date and for what kind of analysis.

5.Which data scientists you admire the most and why?

6.Suppose you are given a data set, what will  you do with it to find out if it suits the business needs of your project or not.

7.What were the business outcomes or decisions for the projects you worked on?

8.What unique skills you think can you add on to our data science team?

9.Which are your favorite data science startups?

10.Why do you want to pursue a career in data science?

11.What have you done to upgrade your skills in analytics?

12.What has been the most useful business insight or development you have found?

13.How will you explain an A/B test to an engineer who does not know statistics?

14.When does parallelism helps your algorithms run faster and when does it make them run slower?

15.How can you ensure that you don’t analyse something that ends up producing meaningless results?

16.How would you explain to the senior management in your organization as to why a particular data set is important?

17.Is more data always better?

18.What are your favourite imputation techniques to handle missing data?

19.What are your favorite data visualization tools?

20.Explain the life cycle of a data science project.

**Data Scientist Interview Questions and Answers**

Data Scientist interview questions asked at a job interview can fall into one of the following categories -

* Technical Data Scientist Interview Questions based on data science programming languages like Python, R, etc.
* Technical Data Scientist Interview Questions based on statistics, probability, math, machine learning, etc.
* Practical experience or Role based data scientist interview questions based on the projects you have worked on, and how they turned out.

**1)Differentiate between Data Science , Machine Learning and AI.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Data Science** | **Machine Learning** | **Artificial Intelligence** |
| Defintion | Data Science is not exactly a subset of machine learning but it uses machine learning to analyse and make future predictions. | A subset of AI that focuses on narrow range of activities. | A wide term that focuses on applications ranging from Robotics to Text Analysis. |
| Role | It can take on a busines role. | It is a purely technical role. | It is a combination of both business and technical aspects. |
| Scope | Data Science is a broad term for diverse disciplines and is not merely about developing and training models. | Machine learning fits within the data science spectrum. | AI is a sub-field of computer science. |
| AI | Loosely integrated | Machine learning is a sub field of AI and is tightly integrated. | A sub- field of computer science consisting of various task like planning, moving around in the world, recognizing objects and sounds, speaking, translating, performing social or business transactions, creative work.. |
| **Data Science vs Machine Learning** | | | |

**2)         Python or R – Which one would you prefer for text analytics?**

The best possible answer for this would be Python because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.

**3)         Which technique is used to predict categorical responses?**

Classification technique is used widely in mining for classifying data sets.

**4)         What is logistic regression? Or State an example when you have used logistic regression recently.**

Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

**5)         What are Recommender Systems?**

A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

**6)         Why data cleaning plays a vital role in 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 take to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.

**7)         Differentiate between univariate, bivariate and multivariate analysis.**

These are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as univariate analysis.

If the analysis attempts to understand the difference between 2 variables at time as in a scatterplot, then it is referred to as bivariate analysis. For example, analysing the volume of sale and a spending can be considered as an example of bivariate analysis.

Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as multivariate analysis.

**8)         What do you understand by the term Normal Distribution?**

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell shaped curve. The random variables are distributed in the form of an symmetrical bell shaped curve.

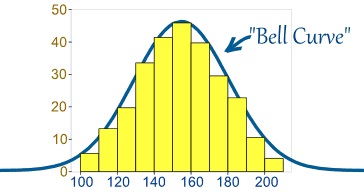


Image Credit: mathisfun.com

**9)         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. X is referred to as the predictor variable and Y as the criterion variable.

**10)       What is 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.

**11)       What is power analysis?**

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

**12)      What is K-means? How can you select K for K-means?**

**13)       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.

**14)       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. Cluster Sample is a probability sample where each sampling unit is a collection, or cluster of elements. Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list,it is progressed from the top again. The best example for systematic sampling is equal probability method.

**15)       Are expected value and mean value different?**

They are not different but the terms are used in different contexts. Mean is generally referred when talking about a probability distribution or sample population whereas expected value is generally referred in a random variable context.

**For Sampling Data**

Mean value is the only value that comes from the sampling data.

Expected Value is the mean of all the means i.e. the value that is built from multiple samples. Expected value is the population mean.

**For Distributions**

Mean value and Expected value are same irrespective of the distribution, under the condition that the distribution is in the same population.

**16)       What does P-value signify about the statistical data?**

P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.

•           P- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.

•           P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.

•           P-value=0.05is the marginal value indicating it is possible to go either way.

**17)  Do gradient descent methods always converge to same point?**

No, they do not because in some cases it reaches a local minima or a local optima point. You don’t reach the global optima point. It depends on the data and starting conditions

**18)  What are categorical variables?**

**19)       A test has a true positive rate of 100% and false positive rate of 5%. There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?**

Let’s suppose you are being tested for a disease, if you have the illness the test will end up saying you have the illness. However, if you don’t have the illness- 5% of the times the test will end up saying you have the illness and 95% of the times the test will give accurate result that you don’t have the illness. Thus there is a 5% error in case you do not have the illness.

Out of 1000 people, 1 person who has the disease will get true positive result.

Out of the remaining 999 people, 5% will also get true positive result.

Close to 50 people will get a true positive result for the disease.

This means that out of 1000 people, 51 people will be tested positive for the disease even though only one person has the illness. There is only a 2% probability of you having the disease even if your reports say that you have the disease.

**20)       How you can make data normal using Box-Cox transformation?**

**21)       What is the difference between Supervised Learning an Unsupervised Learning?**

If an algorithm learns something from the training data so that the knowledge can be applied to the test data, then it is referred to as Supervised Learning. Classification is an example for Supervised Learning. If the algorithm does not learn anything beforehand because there is no response variable or any training data, then it is referred to as unsupervised learning. Clustering is an example for unsupervised learning.

**22) Explain the use of Combinatorics in data science.**

**23) Why is vectorization considered a powerful method for optimizing numerical code?**

**24) What is the goal of A/B Testing?**

It is a statistical hypothesis testing for randomized experiment with two variables A and B. The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of an interest. An example for this could be identifying the click through rate for a banner ad.

**25)       What is an Eigenvalue and Eigenvector?**

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching. Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

**26)       What is Gradient Descent?**

**27)       How can outlier values be treated?**

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values.The most common ways to treat outlier values –

1) To change the value and bring in within a range

2) To just remove the value.

**28)       How can you assess a good logistic model?**

There are various methods to assess the results of a logistic regression analysis-

•           Using Classification Matrix to look at the true negatives and false positives.

•           Concordance that helps identify the ability of the logistic model to differentiate between the event happening and not happening.

•           Lift helps assess the logistic model by comparing it with random selection.

**29)       What are various steps involved in an analytics project?**

•           Understand the business problem

•           Explore the data and become familiar with it.

•           Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.

•           After data preparation, start running the model, analyse the result and tweak the approach. This is an iterative step till the best possible outcome is achieved.

•           Validate the model using a new data set.

•           Start implementing the model and track the result to analyse the performance of the model over the period of time.

**30)** **How can you iterate over a list and also retrieve element indices at the same time?**

This can be done using the enumerate function which takes every element in a sequence just like in a list and adds its location just before it.

**31)       During analysis, how do you treat missing values?**

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored.There are various factors to be considered when answering this question-

* 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.
* Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

**32)       Explain about the box cox transformation in regression models.**

For some reason or the other, the response variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or  follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data  meets the required assumptions. A Box cox transformation is a statistical technique to transform non-mornla dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.

**33)       Can you use machine learning for time series analysis?**

Yes, it can be used but it depends on the applications.

**34)       Write a function that takes in two sorted lists and outputs a sorted list that is their union.**

First solution which will come to your mind is to merge two lists and short them afterwards

**Python code-**  
def return\_union(list\_a, list\_b):  
    return sorted(list\_a + list\_b)

**R code-**  
return\_union <- function(list\_a, list\_b)  
{  
list\_c<-list(c(unlist(list\_a),unlist(list\_b)))  
return(list(list\_c[[1]][order(list\_c[[1]])]))  
}

Generally, the tricky part of the question is not to use any sorting or ordering function. In that case you will have to write your own logic to answer the question and impress your interviewer.

Python code-  
def return\_union(list\_a, list\_b):  
    len1 = len(list\_a)  
    len2 = len(list\_b)  
    final\_sorted\_list = []  
    j = 0  
    k = 0  
  
    for i in range(len1+len2):  
        if k == len1:  
            final\_sorted\_list.extend(list\_b[j:])  
            break  
        elif j == len2:  
            final\_sorted\_list.extend(list\_a[k:])  
            break  
        elif list\_a[k] < list\_b[j]:  
            final\_sorted\_list.append(list\_a[k])  
            k += 1  
        else:  
            final\_sorted\_list.append(list\_b[j])  
            j += 1  
    return final\_sorted\_list

Similar function can be returned in R as well by following the similar steps.

return\_union <- function(list\_a,list\_b)  
{  
#Initializing length variables  
len\_a <- length(list\_a)  
len\_b <- length(list\_b)  
len <- len\_a + len\_b  
  
#initializing counter variables

j=1  
k=1

#Creating an empty list which has length equal to sum of both the lists

list\_c <- list(rep(NA,len))

#Here goes our for loop

for(i in 1:len)  
  {  
    if(j>len\_a)  
      {  
        list\_c[i:len] <- list\_b[k:len\_b]  
        break  
      }  
    else if(k>len\_b)  
      {  
        list\_c[i:len] <- list\_a[j:len\_a]  
        break  
      }  
    else if(list\_a[[j]] <= list\_b[[k]])  
      {  
        list\_c[[i]] <- list\_a[[j]]  
        j <- j+1  
      }  
    else if(list\_a[[j]] > list\_b[[k]])  
    {  
      list\_c[[i]] <- list\_b[[k]]  
      k <- k+1  
    }  
  }  
  return(list(unlist(list\_c)))

  }

**35)       What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)?**

In bayesian estimate we have some knowledge about the data/problem (prior) .There may be several values of the parameters which explain data and hence we can look for multiple parameters like 5 gammas and 5 lambdas that do this. As a result of Bayesian Estimate, we get multiple models for making multiple predcitions i.e. one for each pair of parameters but with the same prior. So, if a new example need to be predicted than computing the weighted sum of these predictions serves the purpose.

Maximum likelihood does not take prior into consideration (ignores the prior) so it is like being a Bayesian  while using some kind of a flat prior.

**36)       What is Regularization and what kind of problems does regularization solve?**

**37)       What is multicollinearity and how you can overcome it?**

**38)        What is the curse of dimensionality?**

**39)        How do you decide whether your linear regression model fits the data?**

**40)       What is the difference between squared error and absolute error?**

**41)       What is Machine Learning?**

The simplest way to answer this question is – we give the data and equation to the machine. Ask the machine to look at the data and identify the coefficient values in an equation.

For example for the linear regression y=mx+c, we give the data for the variable x, y and the machine learns about the values of m and c from the data.

**42) How are confidence intervals constructed and how will you interpret them?**

**43) How will you explain logistic regression to an economist, physican scientist and biologist?**

**44) How can you overcome Overfitting?**

**45) Differentiate between wide and tall data formats?**

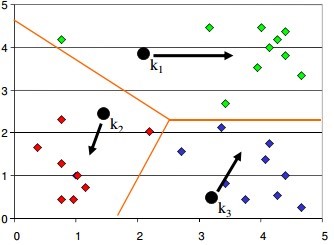
**46) Is Naïve Bayes bad? If yes, under what aspects.**

**47) How would you develop a model to identify plagiarism?**

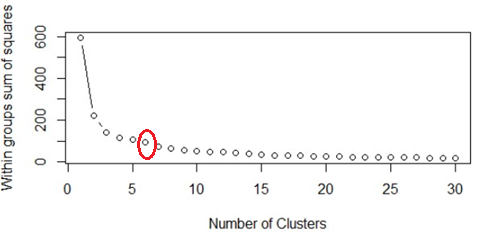
**48) How will you define the number of clusters in a clustering algorithm?**

Though the Clustering Algorithm is not specified, this question will mostly be asked in reference to K-Means clustering where “K” defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.



Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.



Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don’t see any decrement in WSS. This point is known as bending point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendograms and identify the distinct groups from there.

**49) Is it better to have too many false negatives or too many false positives?**

**50) Is it possible to perform logistic regression with Microsoft Excel?**

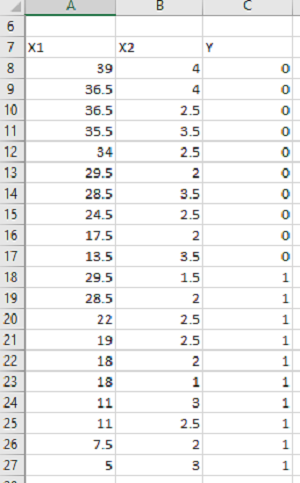
It is possible to perform logistic regression with Microsoft Excel. There are two ways to do it using Excel.

a)      One is to use Add-ins provided by many websites which we can use.

b)      Second is to use fundamentals of logistic regression and use Excel’s computational power to build a logistic regression

But when this question is being asked in an interview, interviewer is not looking for a name of Add-ins rather a method using the base excel functionalities.

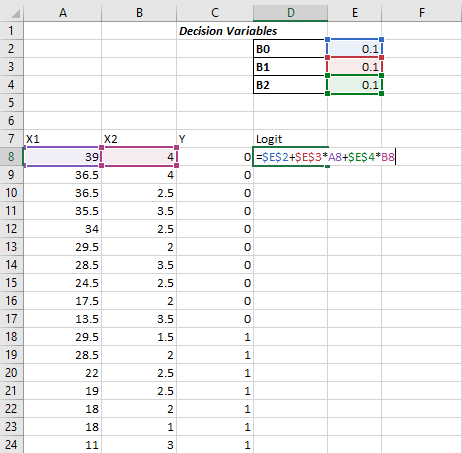
Let’s use a sample data to learn about logistic regression using Excel. (Example assumes that you are familiar with basic concepts of logistic regression)



Data shown above consists of three variables where X1 and X2 are independent variables and Y is a class variable. We have kept only 2 categories for our purpose of binary logistic regression classifier.

Next we have to create a logit function using independent variables, i.e.

Logit = L = β0 +  β1\*X1 +  β2\*X2



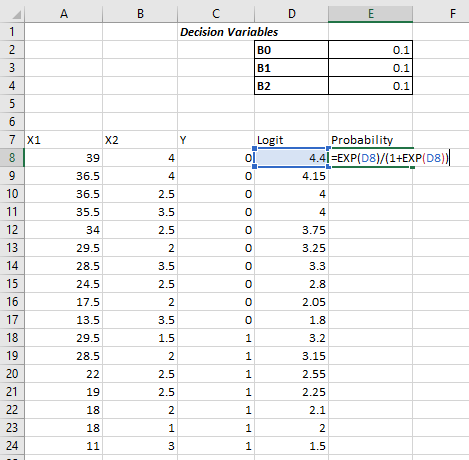
We have kept the initial values of beta 1, beta 2 as 0.1 for now and we will use Excel Solve to optimize the beta values in order to maximize our log likelihood estimate.

Assuming that you are aware of logistic regression basics, we calculate probability values from Logit using following formula:

Probability=  e^Logit/(1+ e^Logit )

e is base of natural logarithm i.e*. e = 2.71828163*

Let’s put it into excel formula to calculate probability values for each of the observation.



The conditional probability   is the probability of Predicted Y, given set of independent variables X.

And this *p*can be calculated as-

P〖(X)〗^Yactual\*[1-P〖(X)〗^(1-Yactual)]

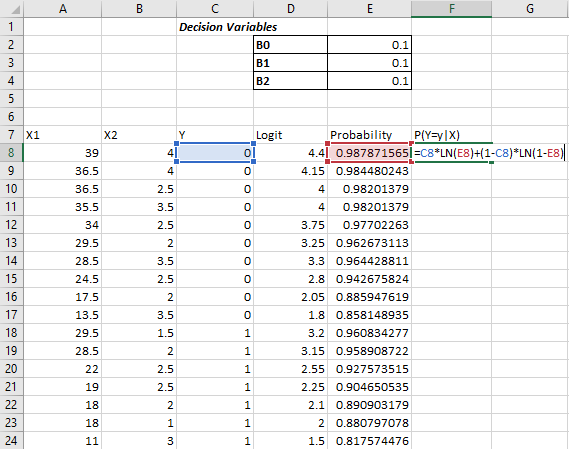
Then we have to take natural log of the above function-

ln⁡〖[ 〗 P〖(X)〗^Yactual\*[1-P(X)^(1-Yactual) ]]

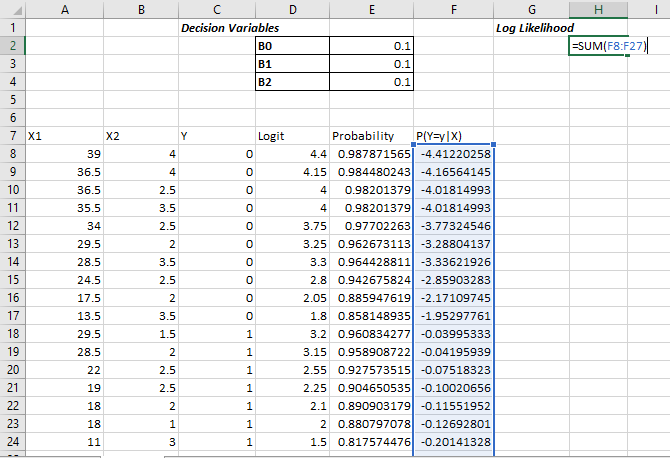
Which turns out to be –

Yactual\*ln⁡〖[ 〗 P(X)]\*(Yactual- 1)\*ln[1-P(X)]

Log likelihood function LL is the sum of above equation for all the observations



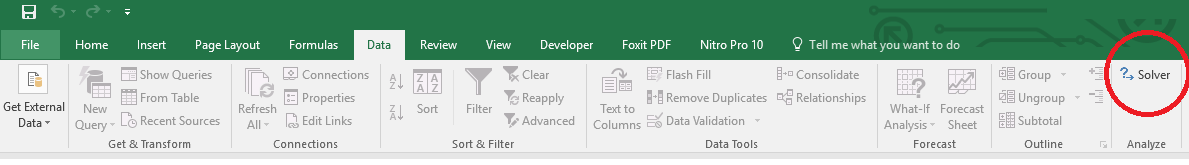
Log likelihood LL will be sum of column G, which we just calculated



The objective is to maximize the Log Likelihood i.e. cell H2 in this example. We have to maximize H2 by optimizing B0, B1, and B2.

We’ll use Excel’s solver add-in to achieve the same.

Excel comes with this Add-in pre-installed and you must see it under Data Tab in Excel as shown below

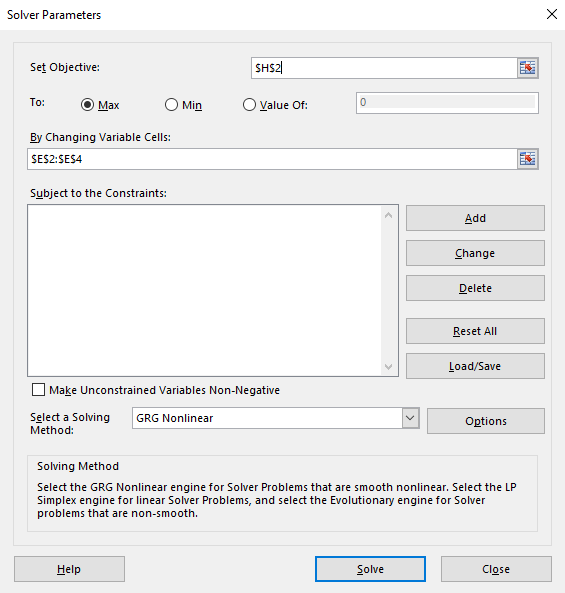


If you don’t see it there then make sure if you have loaded it. To load an add-in in Excel,

Go to *File >> Options >> Add-Ins*and see if checkbox in front of required add-in is checked or not? Make sure to check it to load an add-in into Excel.

If you don’t see Solver Add-in there, go to the bottom of the screen (Manage Add-Ins) and click on OK. Next you will see a popup window which should have your Solver add-in present. Check the checkbox in-front of the add-in name. If you don’t see it there as well click on browse and direct it to the required folder which contains Solver Add-In.

Once you have your Solver loaded, click on Solver icon under Data tab and You will see a new window popped up like –

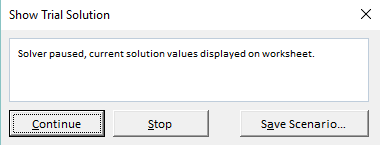


             Put *H2* in set objective, select max and fill cells *E2 to E4* in next form field.

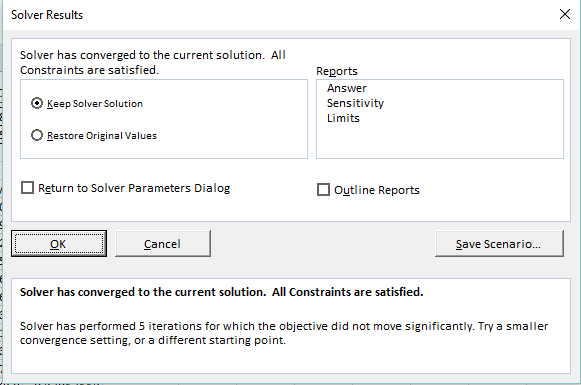
             By doing this we have told Solver to Maximize H2 by changing values in cells E2 to E4.

             Now click on Solve button at the bottom –

            You will see a popup like below -

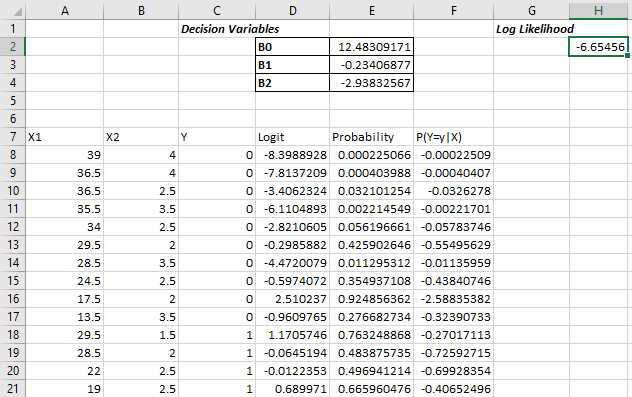


This shows that Solver has found a local maxima solution but we are in need of Global Maxima Output. Keep clicking on Continue until it shows the below popup



It shows that Solver was able to find and converge the solution. In case it is not able to converge it will throw an error. Select “Keep Solver Solution” and Click on OK to accept the solution provided by Solver.

Now, you can see that value of Beta coefficients from B0, B1 B2have changed and our Log Likelihood function has been maximized.



Using these values of Betas you can calculate the probability and hence response variable by deciding the probability cut-off.

**51)  What do you understand by Fuzzy merging ? Which language will you use to handle it?**

**52) What is the difference between skewed and uniform distribution?**

When the observations in a dataset are spread equally across the range of distribution, then it is referred to as uniform distribution. There are no clear perks in an uniform distribution. Distributions that have more observations on one side of the graph than the other  are referred to as skewed distribution.Distributions with fewer observations on the left ( towards lower values) are said to be skewed left and distributions with fewer observation on the right ( towards higher values) are said to be skewed right.

**53) You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?**

Since the question asked, is about post model building exercise, we will assume that you have already tested for null hypothesis, multi collinearity and Standard error of coefficients.

Once you have built the model, you should check for following –

·         Global F-test to see the significance of group of independent variables on dependent variable

·         R^2

·         Adjusted R^2

·         RMSE, MAPE

In addition to above mentioned quantitative metrics you should also check for-

·         Residual plot

·         Assumptions of linear regression

**54) What do you understand by Hypothesis in the content of Machine Learning?**

**55) What do you understand by Recall and Precision?**

Recall  measures "Of all the actual true samples how many did we classify as true?"

Precision measures "Of all the samples we classified as true how many are actually true?"

We will explain this with a simple example for better understanding -

Imagine that your wife gave you surprises every year on your anniversary in last 12 years. One day all of a sudden your wife asks -"Darling, do you remember all anniversary surprises from me?".

This simple question puts your life into danger.To save your life, you need to Recall all 12 anniversary surprises from your memory. Thus, Recall(R) is the ratio of number of events you can correctly recall to the number of all correct events. If you can recall all the 12 surprises correctly then the recall ratio is 1 (100%) but if you can recall only 10 suprises correctly of the 12 then the recall ratio is 0.83 (83.3%).

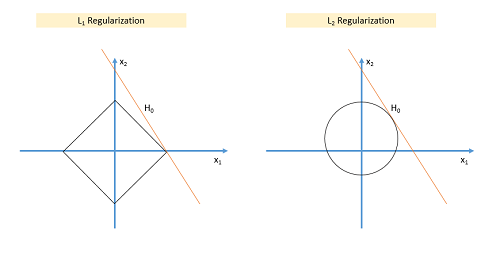
However , you might be wrong in some cases. For instance, you answer 15 times, 10 times the surprises you guess are correct and 5 wrong. This implies that your recall ratio is 100% but the precision is 66.67%.

Precision is the ratio of number of events you can correctly recall to a number of all events you recall (combination of wrong and correct recalls).

**56) How will you find the right K for K-means?**

**57) Why L1 regularizations causes parameter sparsity whereas L2 regularization does not?**

Regularizations in statistics or in the field of machine learning is used to include some extra information in order to solve a problem in a better way. L1 & L2 regularizations are generally used to add constraints to optimization problems.



In the example shown above H0 is a hypothesis. If you observe, in L1 there is a high likelihood to hit the corners as solutions while in L2, it doesn’t. So in L1 variables are penalized more as compared to L2 which results into sparsity.

In other words, errors are squared in L2, so model sees higher error and tries to minimize that squared error.

**58) How can you deal with different types of seasonality in time series modelling?**

Seasonality in time series occurs when time series shows a repeated pattern over time. E.g., stationary sales decreases during holiday season, air conditioner sales increases during the summers etc. are few examples of seasonality in a time series.

Seasonality makes your time series non-stationary because average value of the variables at different time periods. Differentiating a time series is generally known as the best method of removing seasonality from a time series. Seasonal differencing can be defined as a numerical difference between a particular value and a value with a periodic lag (i.e. 12, if monthly seasonality is present)

**59) In experimental design, is it necessary to do randomization? If yes, why?**

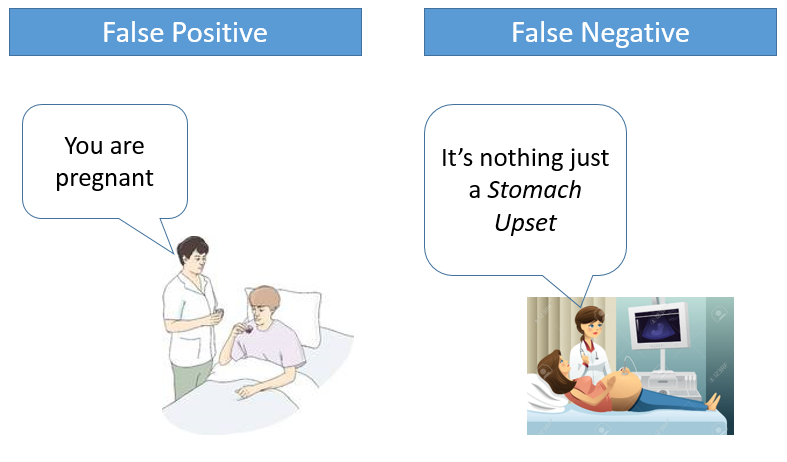
**60) What do you understand by conjugate-prior with respect to Naïve Bayes?**

**61) Can you cite some examples where a false positive is important than a false negative?**

Before we start, let us understand what are false positives and what are false negatives.

False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.

 And, False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.



In medical field, assume you have to give chemo therapy to patients. Your lab tests patients for certain vital information and based on those results they decide to give radiation therapy to a patient.

Assume a patient comes to that hospital and he is tested positive for cancer (But he doesn’t have cancer) based on lab prediction. What will happen to him? (Assuming Sensitivity is 1)

One more example might come from marketing. Let’s say an ecommerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $5000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above 5K.

Now what if they have sent it to false positive cases?

6**2) Can you cite some examples where a false negative important than a false positive?**

Assume there is an airport ‘A’ which has received high security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to shortage of staff they decided to scan passenger being predicted as risk positives by their predictive model.

What will happen if a true threat customer is being flagged as non-threat by airport model?

 Another example can be judicial system. What if Jury or judge decide to make a criminal go free?

 What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after few years and realize that you had a false negative?

**63) Can you cite some examples where both false positive and false negatives are equally important?**

In the banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don’t want to lose good customers and at the same point of time they don’t want to acquire bad customers. In this scenario both the false positives and false negatives be**come very important to measure.**

These days we hear many cases of players using steroids during sport competitions Every player has to go through a steroid test before the game starts. A false positive can ruin the career of a Great sportsman and a false negative can make the game unfair.

**64) Can you explain the difference between a Test Set and a Validation Set?**

Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid Overfitting of the model being built. On the other hand, test set is used for testing or evaluating the performance of a trained machine leaning model.

In simple terms ,the differences can be summarized as-

* Training Set is to fit the parameters i.e. weights.
* Test Set is to assess the performance of the model i.e. evaluating the predictive power and generalization.
* Validation set is to tune the parameters.

**65) What makes a dataset gold standard?**

**66) What do you understand by statistical power of sensitivity and how do you calculate it?**

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, RF etc.). Sensitivity is nothing but “Predicted TRUE events/ Total events”. True events here are the events which were true and model also predicted them as true.

Calculation of senstivity is pretty straight forward-

**Senstivity = True Positives /Positives in Actual Dependent Variable**

Where, True positives are Positive events which are correctly classified as Positives.

**67) What is the importance of having a selection bias?**

Selection Bias occurs when there is no appropriate randomization acheived while selecting individuals, groups or data to be analysed.Selection bias implies that the obtained sample does not exactly represent the population that was actually intended to be analyzed.Selection bias consists of Sampling Bias, Data, Attribute and Time Interval.

**68) Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.**

SVM and Random Forest are both used in classification problems.

a)      If you are sure that your data is outlier free and clean then go for SVM. It is the opposite -   if your data might contain outliers then Random forest would be the best choice

b)      Generally, SVM consumes more computational power than Random Forest, so if you are constrained with memory go for Random Forest [machine learning algorithm](https://www.dezyre.com/article/top-10-machine-learning-algorithms/202).

**c)** Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose Random Forest machine learning algorithm.

d)      Random Forest machine learning algorithms are preferred for multiclass problems.

e)     SVM is preferred in multi-dimensional problem set - like text classification

but as a good data scientist, you should experiment with both of them and test for accuracy or rather you can use ensemble of many Machine Learning techniques.

**69) What do you understand by feature vectors?**

**70) How do data management procedures like missing data handling make selection bias worse?**

Missing value treatment is one of the primary tasks which a data scientist is supposed to do before starting data analysis. There are multiple methods for missing value treatment. If not done properly, it could potentially result into selection bias. Let see few missing value treatment examples and their impact on selection-

**Complete Case Treatment:** Complete case treatment is when you remove entire row in data even if one value is missing. You could achieve a selection bias if your values are not missing at random and they have some pattern. Assume you are conducting a survey and few people didn’t specify their gender. Would you remove all those people? Can’t it tell a different story?

**Available case analysis:**Let say you are trying to calculate correlation matrix for data so you might remove the missing values from variables which are needed for that particular correlation coefficient. In this case your values will not be fully correct as they are coming from population sets.

**Mean Substitution:**In this method missing values are replaced with mean of other available values.This might make your distribution biased e.g., standard deviation, correlation and regression are mostly dependent on the mean value of variables.

Hence, various data management procedures might include selection bias in your data if not chosen correctly.

**71) What are the advantages and disadvantages of using regularization methods like Ridge Regression?**

**72) What do you understand by long and wide data formats?**

**73) What do you understand by outliers and inliers? What would you do if you find them in your dataset?**

**74) Write a program in Python which takes input as the diameter of a coin and weight of the coin and produces output as the money value of the coin.**

**75) What are the basic assumptions to be made for linear regression?**

Normality of error distribution, statistical independence of errors, linearity and additivity.

**76) Can you write the formula to calculat R-square?**

R-Square can be calculated using the below formular -

1 - (Residual Sum of Squares/ Total Sum of Squares)

**77) What is the advantage of performing dimensionality reduction before fitting an SVM?**

Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

**78) How will you assess the statistical significance of an insight whether it is a real insight or just by chance?**

Statistical importance of an insight can be accessed using Hypothesis Testing.

**79) How would you create a taxonomy to identify key customer trends in unstructured data?**

[Tweet: Data Science Interview questions #1 - How would you create a taxonomy to identify key customer trends in unstructured data? - http://ctt.ec/sdqZ0+](http://ctt.ec/sdqZ0)

The best way to approach this question is to mention that it is good to check with the business owner and understand their objectives before categorizing the data. Having done this, it is always good to follow an iterative approach by pulling new data samples and improving the model accordingly by validating it for accuracy by soliciting feedback from the stakeholders of the business. This helps ensure that your model is producing actionable results and improving over the time.

**80) How will you find the correlation between a categorical variable and a continuous variable ?**

You can use the analysis of covariance technqiue to find the correlation between a categorical variable and a continuous variable.

**Here's a list of 20 of the most popular questions you can expect in an interview and how to frame your answers.**

1. What are feature vectors?

Answer:

A feature vector is an n-dimensional vector of numerical features that represent some object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way.

2. Explain the steps in making a decision tree.

Answer:

Take the entire data set as input.

Look for a split that maximizes the separation of the classes. A split is any test that divides the data into two sets.

Apply the split to the input data (divide step).

Re-apply steps 1 to 2 to the divided data.

Stop when you meet some stopping criteria.

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

3. What is root cause analysis?

Answer:

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 reoccurring.

4. What is logistic regression?

Answer:

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

5. What are Recommender Systems?

Answer:

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.

6. Explain cross-validation.

Answer:

It 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 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) in order to limit problems like overfitting and gain insight on how the model will generalize to an independent data set.

7. What is Collaborative Filtering?

Answer:

The process of filtering used by most recommender systems to find patterns and information by collaborating perspectives, numerous data sources, and several agents.

8. Do gradient descent methods at all times converge to a similar point?

Answer:

No, 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.

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

Answer:

This is a 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.

10. What are the drawbacks of the linear model?

Answer:

Some drawbacks of the linear model are:

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

Nervous about your interview? [Enroll in our Data Science course](https://www.simplilearn.com/big-data-and-analytics/data-scientist-certification-sas-r-excel-training?source=GhPreviewCTAText#/course-preview)

15.  What are Eigenvalue and Eigenvector?

Answer:

Eigenvectors are for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

16. Why is resampling done?

Answer:

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)

17. Explain selective bias.

Answer:

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

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

Answer:

Selection bias

Under coverage bias

Survivorship bias

19. Explain survivorship bias.

Answer:

It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

20. How do you work towards a random forest?

Answer:

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

Build several decision trees on bootstrapped training samples of data

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

Rule of thumb: At each split m=p√m=p

Predictions: At the majority rule

For data scientists, the work isn’t easy, but it’s rewarding and there are plenty of available positions out there. Be sure to prepare yourself for the rigors of interviewing and stay sharp with the nuts-and-bolts of data science.

Top Answers to 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.

**R and Python are two of the most important programming languages for Machine Learning Algorithms.**

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:

i.An extensive collection of tools for data analysis

ii.Operators for performing calculations on matrix and array

iii.Data analysis technique for graphical representation

iv. A highly developed yet simple and effective programming language

v. It extensively supports machine learning applications

vi.It acts as a connecting link between various software, tools and datasets

vii. Create high quality reproducible analysis that is flexible and powerful

viii. Provides a robust package ecosystem for diverse needs

ix.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:

i.Determining and analyzing the correlation and direction of the data

ii.Deploying the estimation of the model

iii.Ensuring the usefulness and validity of the model  
iv.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.

**BASIC DATA SCIENCE INTERVIEW QUESTIONS**

1. What is Data Science? Also, list the differences between supervised and unsupervised learning.

Data Science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. By combining aspects of statistics, computer science, applied mathematics, and visualization, data science can turn the vast amounts of data the digital age generates into new insights and new knowledge.

|  |  |
| --- | --- |
| Supervised Learning vs Unsupervised Learning | |
| Supervised Learning | Unsupervised Learning |
| 1. Input data is labeled. | 1. Input data is unlabeled. |
| 2. Uses training dataset. | 2. Uses the input data set. |
| 3. Used for prediction. | 3. Used for analysis. |
| 4. Enables classification and regression. | 4. Enables Classification, Density Estimation, & Dimension Reduction |

2. What are the important skills to have in Python with regard to data analysis?

The following are some of the important skills to possess which will come handy when performing data analysis using Python.

i.Good understanding of the built-in data types especially lists, dictionaries, tuples and sets.

ii.Mastery of N-dimensional NumPy arrays.

iii.Mastery of pandas dataframes.

Iv.Ability to perform element-wise vector and matrix operations on NumPy arrays. This requires the biggest shift in mindset for someone coming from a traditional software development background who’s used to for loops.

V.Knowing that you should use the Anaconda distribution and the conda package manager.

Vi.Familiarity with scikit-learn.

Vii.Ability to write efficient list comprehensions instead of traditional for loops.

Viii.Ability to write small, clean functions (important for any developer), preferably pure functions that don’t alter objects.

Ix.Knowing how to profile the performance of a Python script and how to optimize bottlenecks.

The following will help to tackle any problem in data analytics and machine learning.

3. What is Selection Bias?

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed. It is sometimes referred to as the selection effect. It is the distortion of a statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

The types of selection bias includes:

Sampling bias: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.

Time interval: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.

Data: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.

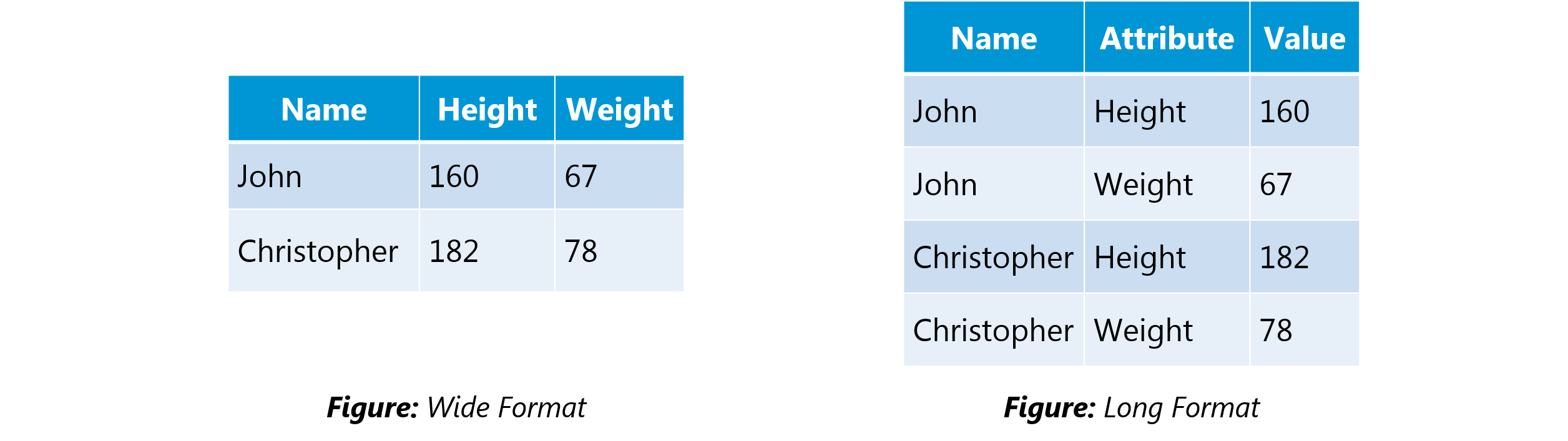
Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

STATISTICS INTERVIEW QUESTIONS

4. What is the difference between “long” and “wide” format data?

In the wide format, a subject’s repeated responses will be in a single row, and each response is in a separate column.

In the long format, each row is a one time point per subject.



5. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up.

However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

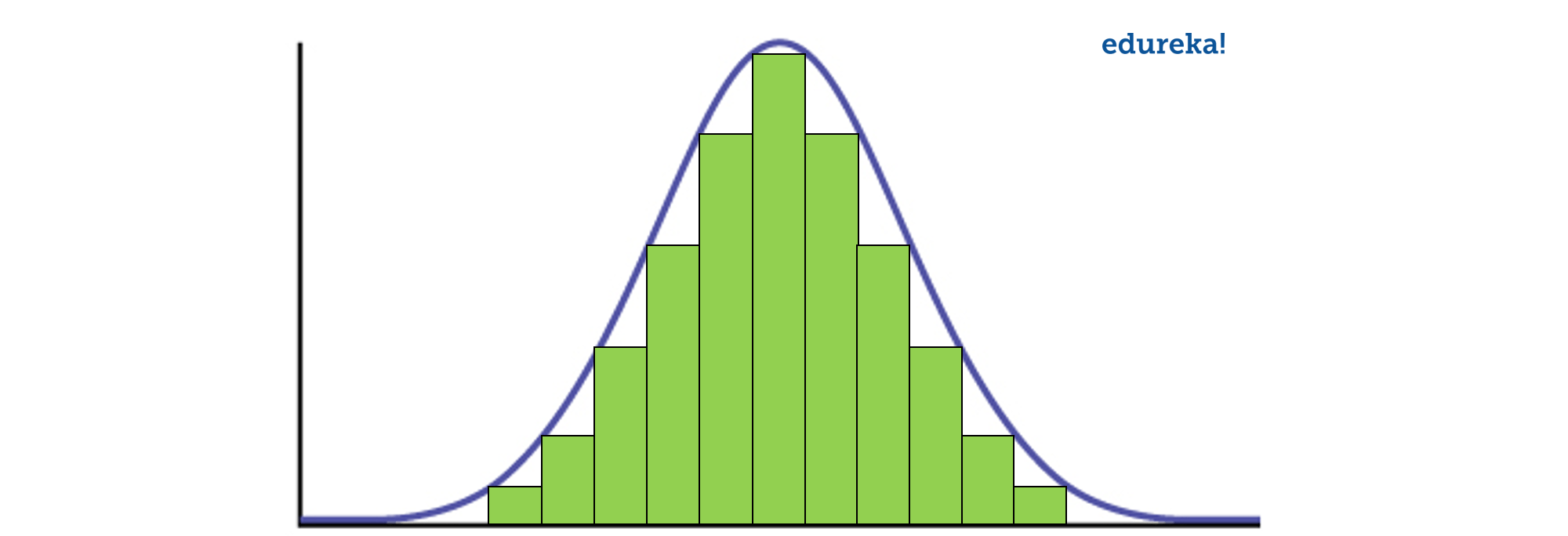


Figure: Normal distribution in a bell curve

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

6. What is the goal of A/B Testing?

It is a statistical hypothesis testing for randomized experiment with two variables A and B.

The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of an interest.

An example for this could be identifying the click-through rate for a banner ad.

7. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).

Sensitivity is nothing but “Predicted True events/ Total events”. True events here are the events which were true and model also predicted them as true.

Calculation of seasonality is pretty straight forward.

Seasonality = ( True Positives ) / ( Positives in Actual Dependent Variable )

where true positives are positive events which are correctly classified as positives.

8. What are the differences between overfitting and underfitting?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.

In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfit has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

DATA ANALYSIS INTERVIEW QUESTIONS

9. Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:

i.Python would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.

Ii.R is more suitable for machine learning than just text analysis.

Iii.Python performs faster for all types of text analytics.

10. How does data cleaning plays a vital role in analysis?

Data cleaning can help in analysis because:

i.Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.

Ii.Data Cleaning helps to increase the accuracy of the model in machine learning.

Iii.It 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 by these sources.

Iv.It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.

11. Differentiate between univariate, bivariate and multivariate analysis.

Univariate analyses are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can the analysis can be referred to as univariate analysis.

Bivariate analysis attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and a spending can be considered as an example of bivariate analysis.

Multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

12. What is Cluster 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. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

13. What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

14. What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

15. Can you cite some examples where a false positive is important than a false negative?

Let us first understand what false positives and false negatives are. False positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error. False negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.

Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn’t have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

Example 2: Let’s say an e-commerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchase.

16. Can you cite some examples where a false negative important than a false positive?

Example 1: Assume there is an airport ‘A’ which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is being flagged as non-threat by airport model?

Example 2: What if Jury or judge decide to make a criminal go free?

Example 3: What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after few years and realize that you had a false negative?

17. Can you cite some examples where both false positive and false negatives are equally important?

In the banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don’t want to lose good customers and at the same point in time, they don’t want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

18. Can you explain the difference between a Validation Set and a Test Set?

Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.

On the other hand, a test set is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

19. 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. Mainly used in backgrounds where the objective is 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) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

20. What is Machine Learning?

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics.

Figure: Applications of Machine Learning

21. What is the Supervised Learning?

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples.

Algorithms: Support Vector Machines, Regression, Naive Bayes, Decision Trees, K-nearest Neighbor Algorithm and Neural Networks

E.g. If you built a fruit classifier, the labels will be “this is an orange, this is an apple and this is a banana”, based on showing the classifier examples of apples, oranges and bananas.

[LEARN DATA SCIENCE TODAY](https://www.edureka.co/data-science)

22. What is Unsupervised learning?

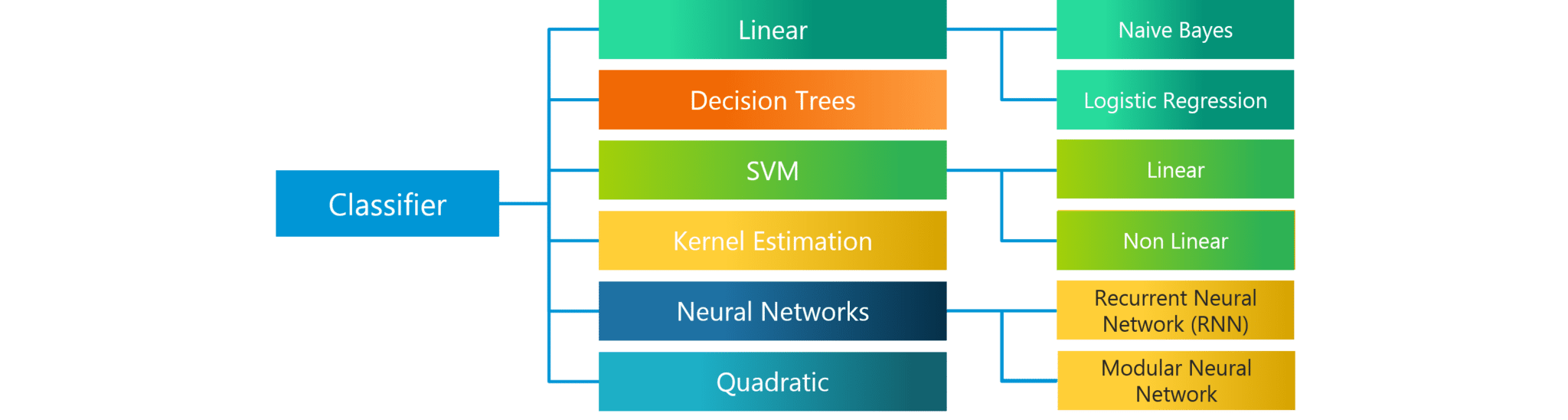
Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

Algorithms: Clustering, Anomaly Detection, Neural Networks and Latent Variable Models

E.g. In the same example, a fruit clustering will categorize as “fruits with soft skin and lots of dimples”, “fruits with shiny hard skin” and “elongated yellow fruits”.

23. What are the various classification algorithms?

The below diagram lists the most important classification algorithms.

Figure: Various Classification algorithms

24. What is logistic regression? State an example when you have used logistic regression recently.

Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables.

For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

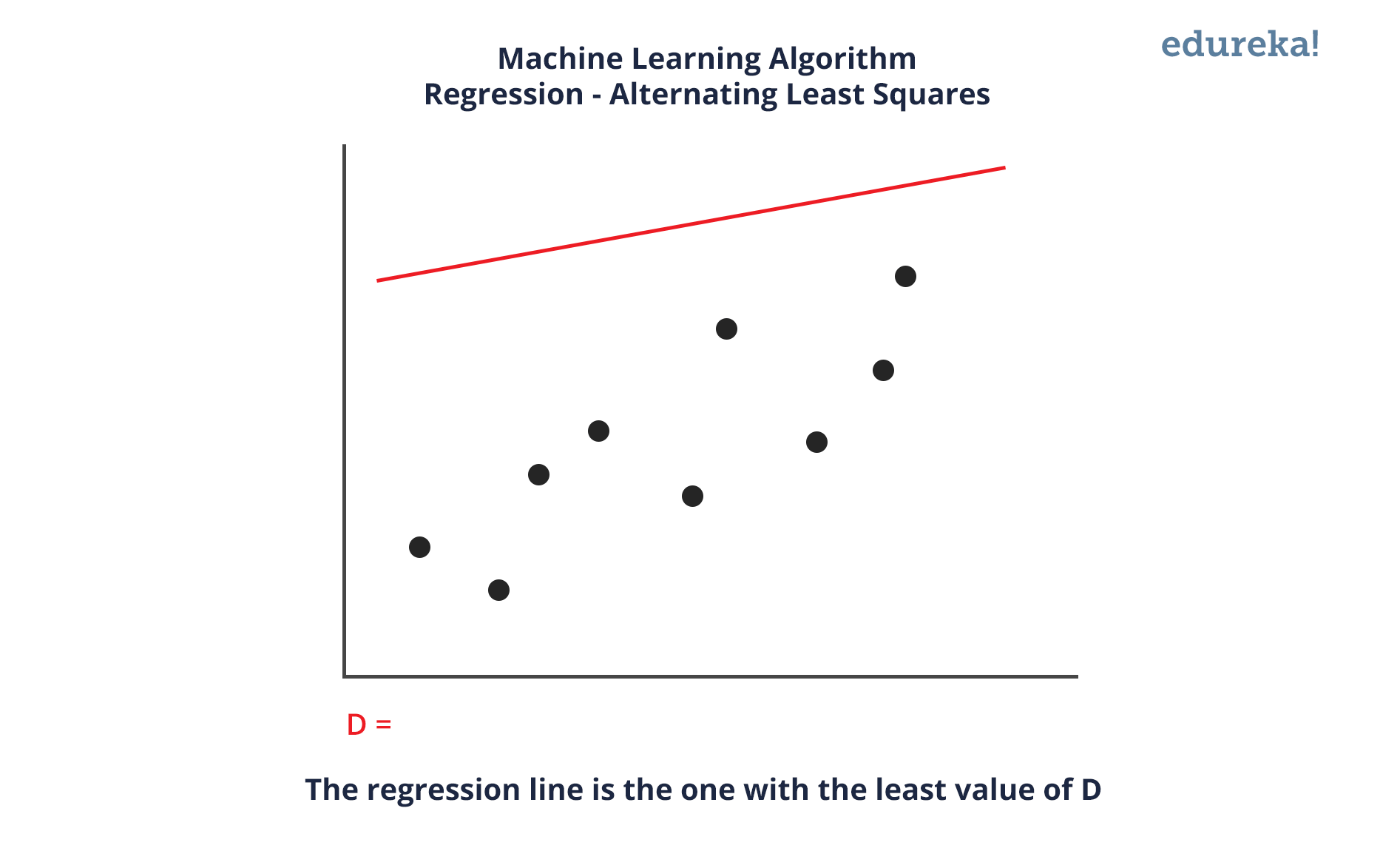
25. What are Recommender Systems?

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. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

Examples include movie recommenders in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

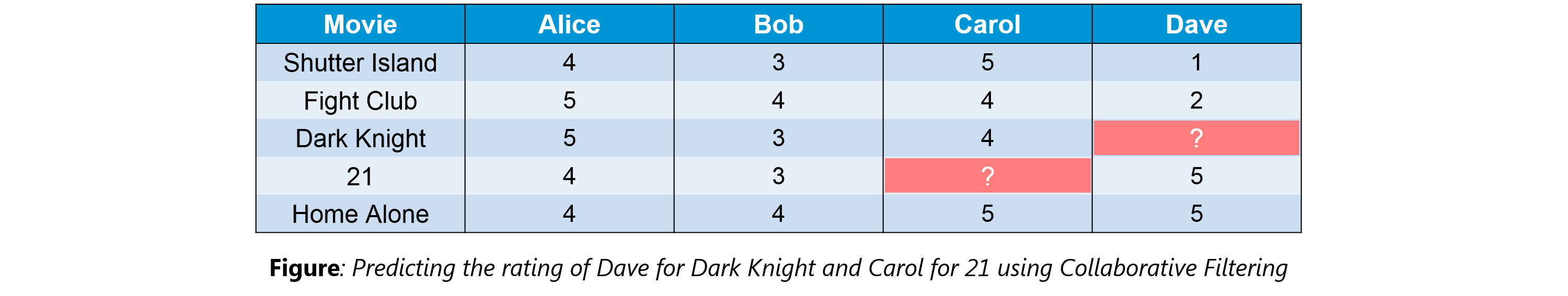
26. What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.



27. 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.



An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others’ ratings for all movies. This concept is widely used in recommending movies in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

28. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values.

All extreme values are not outlier values. The most common ways to treat outlier values

i.To change the value and bring in within a range

ii.To just remove the value.

29. What are various steps involved in an analytics project?

The following are the various steps involved in an analytics project:

i.Understand the business problem

ii.Explore the data and become familiar with it.

Iii.Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.

Iv.After data preparation, start running the model, analyse the result and tweak the approach. This is an iterative step till the best possible outcome is achieved.

5.Validate the model using a new data set.

6.Start implementing the model and track the result to analyse the performance of the model over the period of time.

30. During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

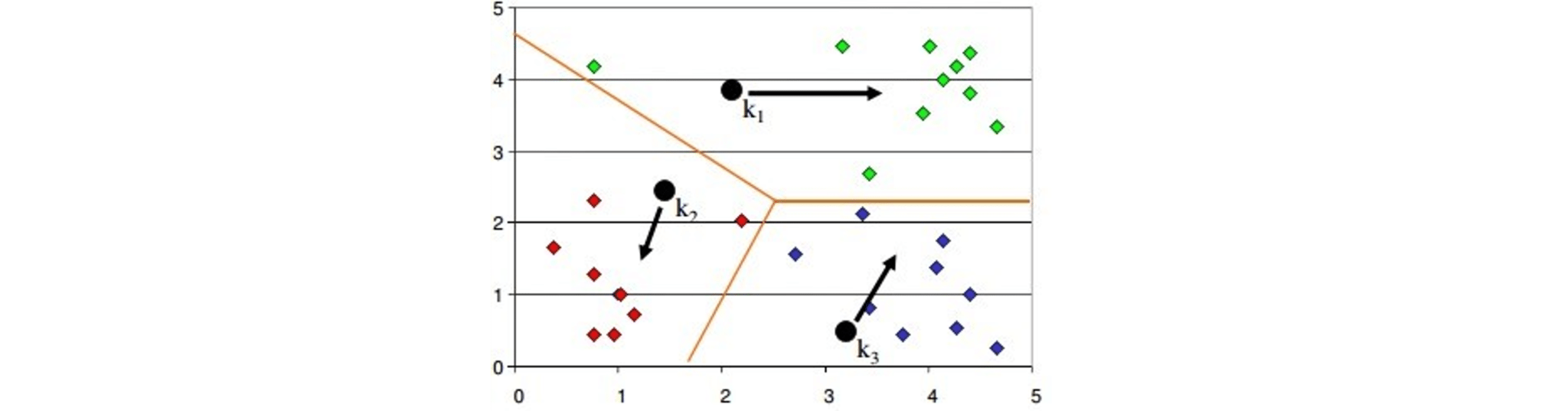
If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

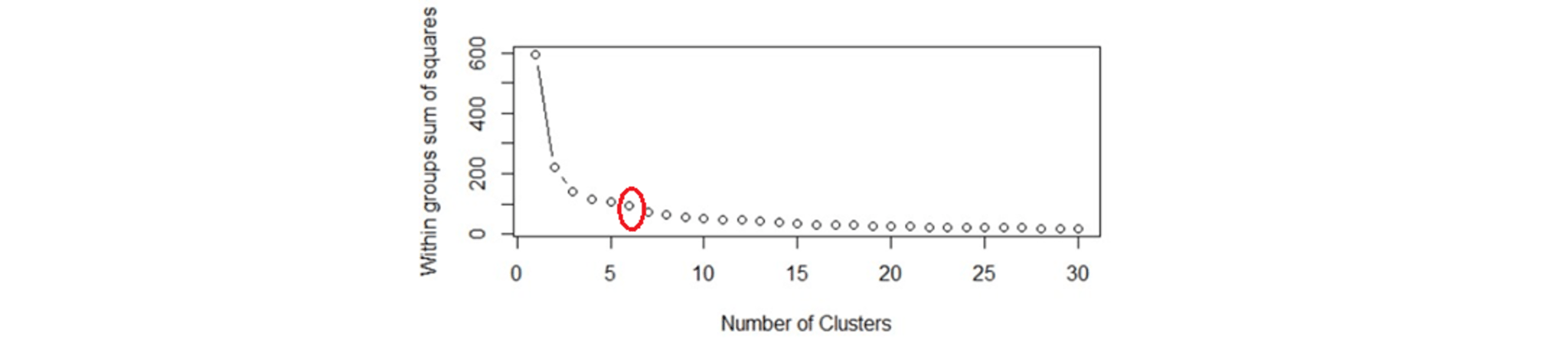
If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

31. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question is mostly in reference to K-Means clustering where “K” defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.

Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below.



1.The Graph is generally known as Elbow Curve.

2.Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don’t see any decrement in WSS.

3.This point is known as bending point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendograms and identify the distinct groups from there.

PROBABILITY INTERVIEW QUESTIONS

32. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the proba­bility that you see at least one shooting star in the period of an hour?

Probability of not seeing any shooting star in 15 minutes is

=   1 – P( Seeing one shooting star )  
=   1 – 0.2          =    0.8

Probability of not seeing any shooting star in the period of one hour

=   (0.8) ^ 4        =    0.4096

Probability of seeing at least one shooting star in the one hour

=   1 – P( Not seeing any star )  
=   1 – 0.4096     =    0.5904

33. How can you generate a random number between 1 – 7 with only a die?

1.Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.

2.To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.

3.A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.

4.All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

34. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

In the case of two children, there are 4 equally likely possibilities

BB, BG, GB and GG;

where B = Boy and G = Girl and the first letter denotes the first child.

From the question, we can exclude the first case of BB. Thus from the remaining 3 possibilities of BG, GB &BB, we have to find the probability of the case with two girls.

Thus, P(Having two girls given one girl)   =    1 / 3

35. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

Probability of selecting fair coin = 999/1000 = 0.999  
Probability of selecting unfair coin = 1/1000 = 0.001

Selecting 10 heads in a row = Selecting fair coin \* Getting 10 heads  +  Selecting an unfair coin

P (A)  =  0.999 \* (1/2)^5  =  0.999 \* (1/1024)  =  0.000976  
P (B)  =  0.001 \* 1  =  0.001  
P( A / A + B )  = 0.000976 /  (0.000976 + 0.001)  =  0.4939  
P( B / A + B )  = 0.001 / 0.001976  =  0.5061

Probability of selecting another head = P(A/A+B) \* 0.5 + P(B/A+B) \* 1 = 0.4939 \* 0.5 + 0.5061  =  0.7531

We’ve broken the interview questions for data scientists into six different categories: statistics, programming, modeling, behavior, culture, and problem-solving.

1.[Statistics](https://www.springboard.com/blog/data-science-interview-questions/#statistics)

2.[Programming](https://www.springboard.com/blog/data-science-interview-questions/#programming)

i.General

ii.Big Data

iii.Python

iv.R

v.SQL

3.[Modeling](https://www.springboard.com/blog/data-science-interview-questions/#modeling)

4.[Behavioral](https://www.springboard.com/blog/data-science-interview-questions/#behavioral)

5.[Culture Fit](https://www.springboard.com/blog/data-science-interview-questions/#culture)

6.[Problem-Solving](https://www.springboard.com/blog/data-science-interview-questions/#problem)

1. Statistics Interview Questions

Statistical computing is the process through which data scientists take raw data and create predictions and models. Without an advanced knowledge of statistics it is difficult to succeed as a data scientist–accordingly, it is likely a good interviewer will try to probe your understanding of the subject matter with statistics-oriented data science interview questions. Be prepared to answer some fundamental statistics questions as part of your data science interview.

Here are examples of rudimentary statistics questions we’ve found:

What is the Central Limit Theorem and why is it important?

“Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can’t obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly.” Read more [here](https://spin.atomicobject.com/2015/02/12/central-limit-theorem-intro/).

What is sampling? How many sampling methods do you know?

“Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.” Read the full answer [here](https://searchbusinessanalytics.techtarget.com/definition/data-sampling).

What is the difference between type I vs type II error?

“A type I error occurs when the null hypothesis is true, but is rejected. A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected.” Read the full answer [here](https://www.datasciencecentral.com/profiles/blogs/understanding-type-i-and-type-ii-errors).

What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?

A linear regression is a good tool for quick predictive analysis: for example, the price of a house depends on a myriad of factors, such as its size or its location. In order to see the relationship between these variables, we need to build a linear regression, which predicts the line of best fit between them and can help conclude whether or not these two factors have a positive or negative relationship. Read more [here](https://www.springboard.com/blog/linear-regression-in-python-a-tutorial/) and [here.](http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients)

What are the assumptions required for linear regression?

There are four major assumptions: 1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data, 2. The errors or residuals of the data are normally distributed and independent from each other, 3. There is minimal multicollinearity between explanatory variables, and 4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

What is a statistical interaction?

”Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another factor.” Read more [here](http://icbseverywhere.com/blog/mini-lessons-tutorials-and-support-pages/statistical-interactions/).

What is selection bias?

“Selection (or ‘sampling’) bias occurs in an ‘active,’ sense when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see. That is, active selection bias occurs when a subset of the data are systematically (i.e., non-randomly) excluded from analysis.” Read more [here](https://www.elderresearch.com/blog/selection-bias-in-analytics).

What is an example of a data set with a non-Gaussian distribution?

“The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate.” Read more [here](https://www.quora.com/Most-machine-learning-datasets-are-in-Gaussian-distribution-Where-can-we-find-the-dataset-which-follows-Bernoulli-Poisson-gamma-beta-etc-distribution).

What is the Binomial Probability Formula?

“The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of π (the Greek letter pi) of occurring.” Read more [here](http://onlinestatbook.com/2/probability/binomial.html).

**2. Programming**

To test your programming skills, employers will typically include two specific data science interview questions: they’ll ask how you would solve programming problems in theory without writing out the code, and then they will also offer whiteboarding exercises for you to code on the spot. For the latter types of questions, we will provide a few examples below, but if you’re looking for in-depth practice solving coding challenges, visit [HackerRank](https://www.hackerrank.com/" \t "_blank). With a “learn by doing” philosophy, there are challenges organized around core concepts commonly tested during interviews.

**2.1 General**

1.With which programming languages and environments are you most comfortable working?

2.What are some pros and cons about your favorite statistical software?

3.Tell me about an original algorithm you’ve created.

4.Describe a data science project in which you worked with a substantial programming component. What did you learn from that experience?

5.Do you contribute to any open-source projects?

6.How would you clean a data set in (insert language here)?

7.Tell me about the coding you did during your last project?

**2.2 Big Data**

1.What are two main components of the Hadoop framework?

The Hadoop Distributed File System (HDFS), MapReduce, and YARN. Read more[here](https://www.quora.com/What-are-the-main-components-of-a-Hadoop-Application).

2.Explain how MapReduce works as simply as possible.

“MapReduce is a programming model that enables distributed processing of large data sets on compute clusters of commodity hardware. Hadoop MapReduce first performs mapping which involves splitting a large file into pieces to make another set of data.” Read more [here](https://bigdata-madesimple.com/basic-components-of-hadoop-architecture-frameworks-used-for-data-science/).

3.How would you sort a large list of numbers?

4.Say you’re given a large data set. What would be your plan for dealing with outliers? How about missing values? How about transformations?

**2.3 Python**

1.What modules/libraries are you most familiar with? What do you like or dislike about them?

2.In Python, how is memory managed?

In Python, memory is managed in a private heap space. This means that all the objects and data structures will be located in a private heap. However, the programmer won’t be allowed to access this heap. Instead, the Python interpreter will handle it. At the same time, the core API will enable access to some Python tools for the programmer to start coding. The memory manager will allocate the heap space for the Python objects while the inbuilt garbage collector will recycle all the memory that’s not being used to boost available heap space. Read more [here](https://www.springboard.com/blog/python-interview-questions/).

3.What are the supported data types in Python?

“Python’s built-in (or standard) data types can be grouped into several classes. Sticking to the hierarchy scheme used in the official Python documentation these are numeric types, sequences, sets and mappings.” Read more [here](https://www.quora.com/What-are-the-supported-data-types-in-Python).

4.What is the difference between a tuple and a list in Python?

“Apart from tuples being immutable there is also a semantic distinction that should guide their usage.” Read more [here](https://stackoverflow.com/questions/626759/whats-the-difference-between-lists-and-tuples).

**2.5 SQL**  
  
Often, SQL questions are case-based, meaning that an employer will task you with solving an SQL problem in order to test your skills from a practical standpoint. For example, you could be given a table and asked to extract relevant data, then filter and order the data as you see fit, and finally report your findings. If you do not feel ready to do this in an interview setting, [Mode Analytics](https://community.modeanalytics.com/sql/tutorial/introduction-to-sql/) has a delightful introduction to [using SQL](https://www.springboard.com/blog/sql-what-you-need-to-know/) that will teach you these commands through an interactive SQL environment.

1.What is the purpose of the group functions in SQL? Give some examples of group functions.

Group functions are necessary to get summary statistics of a data set. COUNT, MAX, MIN, AVG, SUM, and DISTINCT are all group functions.

2.Tell me the difference between an inner join, left join/right join, and union.

“In a Venn diagram the inner join is when both tables have a match, a left join is when there is a match in the left table and the right table is null, a right join is the opposite of a left join, and a full join is all of the data combined.” Read more [here](https://www.quora.com/SQL-What-is-the-difference-between-inner-join-left-join-right-join-and-full-join).

3.What does UNION do? What is the difference between UNION and UNION ALL?

“UNION removes duplicate records (where all columns in the results are the same), UNION ALL does not.” Read more [here](https://stackoverflow.com/questions/49925/what-is-the-difference-between-union-and-union-all).

4.What is the difference between SQL and MySQL or SQL Server?

“SQL stands for Structured Query Language. It’s a standard language for accessing and manipulating databases. MySQL is a database management system, like SQL Server, Oracle, Informix, Postgres, etc.” Read more [here](https://www.quora.com/What-is-the-difference-between-SQL-and-MySQL-or-SQL-Server).

5.If a table contains duplicate rows, does a query result display the duplicate values by default? How can you eliminate duplicate rows from a query result?

Yes. One way you can eliminate duplicate rows with the DISTINCT clause.

3. Modeling

Data modeling is where a data scientist provides value for a company. Turning data into predictive and actionable information is difficult, talking about it to a potential employer even more so. Practice describing your past experiences building models–what were the techniques used, challenges overcome, and successes achieved in the process? The group of questions below are designed to uncover that information, as well as your formal education of different modeling techniques. If you can’t describe the theory and assumptions associated with a model you’ve used, it won’t leave a good impression.

Take a look at the questions below to practice. Not all of the questions will be relevant to your interview–you’re not expected to be a master of all techniques. The best use of these questions is to re-familiarize yourself with the modeling techniques you’ve learned in the past.

1.Tell me about how you designed a model for a past employer or client.

2.What are your favorite data visualization techniques?

3.How would you effectively represent data with 5 dimensions?

4.How is k-NN different from k-means clustering?

k-NN, or k-nearest neighbors is a classification algorithm, where the k is an integer describing the number of neighboring data points that influence the classification of a given observation. K-means is a clustering algorithm, where the k is an integer describing the number of clusters to be created from the given data.

5.How would you create a logistic regression model?

6.Have you used a time series model? Do you understand cross-correlations with time lags?

7.Explain the 80/20 rule, and tell me about its importance in model validation.

“People usually tend to start with a 80-20% split (80% training set – 20% test set) and split the training set once more into a 80-20% ratio to create the validation set.” Read more [here](https://www.beyondthelines.net/machine-learning/how-to-split-a-dataset/).

8.Explain what precision and recall are. How do they relate to the ROC curve?

Recall describes what percentage of true positives are described as positive by the model. Precision describes what percent of positive predictions were correct. The ROC curve shows the relationship between model recall and specificity–specificity being a measure of the percent of true negatives being described as negative by the model. Recall, precision, and the ROC are measures used to identify how useful a given classification model is. Read more [here](http://www.kdnuggets.com/faq/precision-recall.html).

9.Explain the difference between L1 and L2 regularization methods.

“A regression model that uses L1 regularization technique is called Lasso Regression and model which uses L2 is called Ridge Regression. The key difference between these two is the penalty term.” Read more [here](https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c).

10.What is root cause analysis?

“All of us dread that meeting where the boss asks ‘why is revenue down?’ The only thing worse than that question is not having any answers! There are many changes happening in your business every day, and often you will want to understand exactly what is driving a given change — especially if it is unexpected. Understanding the underlying causes of change is known as root cause analysis.” Read more [here](https://towardsdatascience.com/how-to-conduct-a-proper-root-cause-analysis-789b9847f84b).

11.What are hash table collisions?

“If the range of key values is larger than the size of our hash table, which is usually always the case, then we must account for the possibility that two different records with two different keys can hash to the same table index. There are a few different ways to resolve this issue. In hash table vernacular, this solution implemented is referred to as collision resolution.” Read more [here](https://medium.com/@bartobri/hash-crash-the-basics-of-hash-tables-bef82a8ea550).

12 What is an exact test?

“In statistics, an exact (significance) test is a test where all assumptions, upon which the derivation of the distribution of the test statistic is based, are met as opposed to an approximate test (in which the approximation may be made as close as desired by making the sample size big enough). This will result in a significance test that will have a false rejection rate always equal to the significance level of the test. For example an exact test at significance level 5% will in the long run reject true null hypotheses exactly 5% of the time.” Read more[here](https://en.wikipedia.org/wiki/Exact_test).

13 In your opinion, which is more important when designing a machine learning model: model performance or model accuracy?

Here’s [one approach to this question](https://en.wikipedia.org/wiki/Accuracy_paradox).

14 What is one way that you would handle an imbalanced data set that’s being used for prediction (i.e., vastly more negative classes than positive classes)?

15 How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression?

16 I have two models of comparable accuracy and computational performance. Which one should I choose for production and why?

17 How do you deal with sparsity?

18 Is it better to spend five days developing a 90-percent accurate solution or 10 days for 100-percent accuracy?

19 What are some situations where a general linear model fails?

Read about this [here](https://www.quora.com/What-are-the-limitations-of-linear-regression-modeling-in-data-analysis).

20 Do you think 50 small decision trees are better than a large one? Why?

Read about this [here](https://www.quora.com/Do-you-think-50-small-decision-trees-are-better-than-a-large-one-Why).

21 When modifying an algorithm, how do you know that your changes are an improvement over not doing anything?

22 Is it better to have too many false positives or too many false negatives?

It depends on several factors. Read about this [here](https://www.quora.com/In-Data-Science-is-it-preferable-to-have-too-many-false-negatives-or-too-many-false-positives).

**4. Past Behavior**

Employers love behavioral questions. They reveal information about the work experience of the interviewee and about their demeanor and how that could affect the rest of the team. From these questions, an interviewer wants to see how a candidate has reacted to situations in the past, how well they can articulate what their role was, and what they learned from their experience.

There are several categories of behavioral questions you’ll be asked:

1 Teamwork

2 Leadership

3 Conflict management

4 Problem-solving

**5 Failure**

Before the interview, write down examples of work experiences related to these topics to refresh your memory—you will need to recall specific examples to answer the questions well. When asked about a prior experience, make sure you tell a story. Being able to concisely and logically craft a story to detail your experiences is important. For example: ”I was asked X, I did A, B, and C, and decided that the answer was Y.”

Of course, if you can highlight experiences having to do with data science, these questions present a great opportunity to showcase a unique accomplishment as a data scientist that you may not have discussed previously.

Here are examples of these sorts of questions/prompts:

1 Tell me about a time when you took initiative.

2 Tell me about a time when you had to overcome a dilemma.

3 Tell me about a time when you resolved a conflict.

4 Tell me about a time you failed and what you have learned from it.

5 Tell me about (a job on your resume). Why did you choose to do it and what do you like most about it?

6 Tell me about a challenge you have overcome while working on a group project.

7 When you encountered a tedious, boring task, how would you deal with it and motivate yourself to complete it?

8 What have you done in the past to make a client satisfied/happy?

9 What have you done in your previous job that you are really proud of?

10 What do you do when your personal life is running over into your work life?

**5. Culture Fit**

If an employer asks you a question on this list, they are trying to get a sense of who you are and how you would fit with the company. They’re trying to gauge where your interest in data science and in the hiring company come from. Take a look at these examples and think about what your best answer would be, but keep in mind that it’s important to be honest with these answers. There’s no reason to not be yourself. There are no right answers to these questions, but the best answers are communicated with confidence.

1 Which data scientists do you admire most? Which startups?

There are plenty of amazing data scientists to choose from—take a look at [this article](https://www.springboard.com/blog/top-data-scientists-on-twitter/) on top data science influencers for interesting information about some of the top data scientists in the world.

2 What do you think makes a good data scientist?

3 How did you become interested in data science?

4 Give a few examples of “best practices” in data science.

5 What is the latest data science book / article you read? What is the latest data mining conference / webinar / class / workshop / training you attended?

If you haven’t read a good data science book recently, Springboard compiled [a list of the best data science books to read](https://www.springboard.com/blog/eleven-best-data-science-books/)! And check out these [data science podcasts](https://www.springboard.com/blog/15-data-science-podcasts/).

6 What’s a project you would want to work on at our company?

7 What unique skills do you think you’d bring to the team?

8 What data would you love to acquire if there were no limitations?

9 Have you ever thought about creating your own startup? Around which idea / concept?

10 What can your hobbies tell me that your [resume](https://www.springboard.com/blog/data-science-resume/) can’t?

11 What are your top 5 predictions for the next 20 years?

12 What did you do today? Or what did you do this week / last week?

13 If you won a million dollars in the lottery, what would you do with the money?

14 What is one thing you believe that most people do not?

15 What personality traits do you butt heads with?

16 What (outside of data science) are you passionate about?

**6. Problem-Solving**

Interviewers will, at some point during the interview process, want to test your problem-solving ability through data science interview questions. Often these tests will be presented as an open-ended question: How would you do X? In general, that X will be a task or problem specific to the company you are applying with. For example, an interviewer at Yelp may ask a candidate how they would create [a system to detect fake Yelp reviews](https://www.springboard.com/blog/eat-rate-love-an-exploration-of-r-yelp-and-the-search-for-good-indian-food/).

Some quick tips: Don’t be afraid to ask questions. Employers want to test your critical thinking skills—and asking questions that clarify points of uncertainty is a trait that any data scientist should have. Also, if the problem offers an opportunity to show off your white-board coding skills or to create schematic diagrams—use that to your advantage. It shows technical skill, and helps to communicate your thought process through a different mode of communication. Always share your thought process—process is often more important than the results themselves for the interviewer.

1 How would you come up with a solution to identify plagiarism?

2 How many “useful” votes will a Yelp review receive?

3 How do you detect individual paid accounts shared by multiple users?

4 You are about to send a million emails. How do you optimize delivery? How do you optimize response?

5 You have a data set containing 100,000 rows and 100 columns, with one of those columns being our dependent variable for a problem we’d like to solve. How can we quickly identify which columns will be helpful in predicting the dependent variable. Identify two techniques and explain them to me as though I were 5 years old.

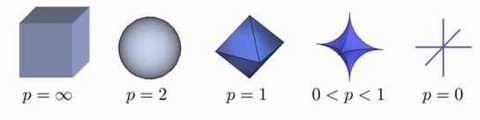
6 How would you detect bogus reviews, or bogus Facebook accounts used for bad purposes?

This is an opportunity to showcase your knowledge of machine learning algorithms; specifically, sentiment analysis and text analysis algorithms. Showcase your knowledge of fraudulent behavior—[what are the abnormal behaviors](https://www.cs.uic.edu/~liub/FBS/fake-reviews.html) that can typically be seen from fraudulent accounts?

7 How would you perform clustering on a million unique keywords, assuming you have 10 million data points—each one consisting of two keywords, and a metric measuring how similar these two keywords are? How would you create this 10 million data points table in the first place?

8 How would you optimize a web crawler to run much faster, extract better information, and better summarize data to produce cleaner databases?

Q1. Explain what regularization is and why it is useful.

Answer by [Matthew Mayo.](https://www.kdnuggets.com/author/matt-mayo)   
  
Regularization is the process of adding a tuning parameter to a model to induce smoothness in order to prevent [overfitting](https://en.wikipedia.org/wiki/Overfitting). (see also KDnuggets posts on [Overfitting](https://www.kdnuggets.com/tag/overfitting))   
  
  
This is most often done by adding a constant multiple to an existing weight vector. This constant is often either the [L1 (Lasso)](https://en.wikipedia.org/wiki/Lasso_(statistics)) or [L2 (ridge)](https://en.wikipedia.org/wiki/Tikhonov_regularization), but can in actuality can be any norm. The model predictions should then minimize the mean of the loss function calculated on the regularized training set.   
  
Xavier Amatriain presents a [good comparison of L1 and L2 regularization here](https://www.quora.com/What-is-the-difference-between-L1-and-L2-regularization), for those interested.    
Fig 1: Lp ball: As the value of p decreases, the size of the corresponding L-p space also decreases. 

Q2. Which data scientists do you admire most? which startups?

Answer by [Gregory Piatetsky](https://www.kdnuggets.com/author/gregory-piatetsky):   
  
This question does not have a correct answer, but here is my personal list of 12 Data Scientists I most admire, not in any particular order.   
  
   
  
[Geoff Hinton](https://www.kdnuggets.com/tag/geoff-hinton), [Yann LeCun](https://www.kdnuggets.com/tag/yann-lecun), and [Yoshua Bengio](https://www.kdnuggets.com/tag/yoshua-bengio" \t "_blank) - for persevering with Neural Nets when and starting the current Deep Learning revolution.   
  
Demis Hassabis, for his amazing work on [DeepMind](https://www.kdnuggets.com/tag/deepmind), which achieved human or superhuman performance on Atari games and recently [Go](https://www.kdnuggets.com/2016/02/google-great-gains-game-go.html).   
  
[Jake Porway](https://www.kdnuggets.com/tag/jake-porway) from DataKind and [Rayid Ghani](https://www.kdnuggets.com/tag/rayid-ghani" \t "_blank) from U. Chicago/DSSG, for enabling data science contributions to social good.   
  
[DJ Patil](https://www.kdnuggets.com/tag/dj-patil), First US Chief Data Scientist, for using Data Science to make US government work better.   
  
[Kirk D. Borne](https://www.kdnuggets.com/tag/kirk-d-borne) for his influence and leadership on social media.   
  
[Claudia Perlich](https://www.kdnuggets.com/tag/claudia-perlich) for brilliant work on ad ecosystem and serving as a great KDD-2014 chair.   
  
[Hilary Mason](https://www.kdnuggets.com/tag/hilary-mason) for great work at Bitly and inspiring others as a Big Data Rock Star.   
  
[Usama Fayyad](https://www.kdnuggets.com/tag/usama-fayyad), for showing leadership and setting high goals for KDD and Data Science, which helped inspire me and many thousands of others to do their best.   
  
[Hadley Wickham](https://www.kdnuggets.com/?s=Hadley+Wickham), for his fantastic work on Data Science and Data Visualization in R, including dplyr, ggplot2, and Rstudio.   
  
There are too many excellent startups in Data Science area, but I will not list them here to avoid a conflict of interest.   
  
Here is some of our previous coverage of [startups](https://www.kdnuggets.com/tag/startups). 

Q3. How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression.

Answer by [Matthew Mayo.](https://www.kdnuggets.com/author/matt-mayo)   
  
[Proposed methods](http://support.sas.com/resources/papers/proceedings12/333-2012.pdf) for model validation: 

1 If the values predicted by the model are far outside of the response variable range, this would immediately indicate poor estimation or model inaccuracy.

2 If the values seem to be reasonable, examine the parameters; any of the following would indicate poor estimation or multi-collinearity: opposite signs of expectations, unusually large or small values, or observed inconsistency when the model is fed new data.

3 Use the model for prediction by feeding it new data, and use the [coefficient of determination](https://en.wikipedia.org/wiki/Coefficient_of_determination) (R squared) as a model validity measure.

4 Use data splitting to form a separate dataset for estimating model parameters, and another for validating predictions.

5 Use [jackknife resampling](https://en.wikipedia.org/wiki/Jackknife_resampling) if the dataset contains a small number of instances, and measure validity with R squared and [mean squared error](https://en.wikipedia.org/wiki/Mean_squared_error) (MSE).

Q4. Explain what precision and recall are. How do they relate to the ROC curve?

Answer by [Gregory Piatetsky](https://www.kdnuggets.com/author/gregory-piatetsky):   
  
Here is the answer from [KDnuggets FAQ: Precision and Recall](https://www.kdnuggets.com/faq/precision-recall.html" \t "_blank): 

Calculating precision and recall is actually quite easy. Imagine there are 100 positive cases among 10,000 cases. You want to predict which ones are positive, and you pick 200 to have a better chance of catching many of the 100 positive cases.  You record the IDs of your predictions, and when you get the actual results you sum up how many times you were right or wrong. There are four ways of being right or wrong:

1 TN / True Negative: case was negative and predicted negative

3 TP / True Positive: case was positive and predicted positive

3 FN / False Negative: case was positive but predicted negative

4 FP / False Positive: case was negative but predicted positive

Makes sense so far? Now you count how many of the 10,000 cases fall in each bucket, say:

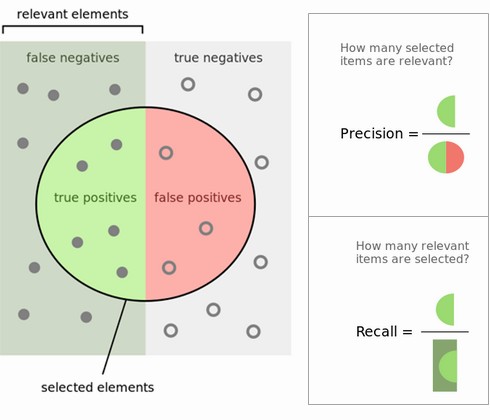
|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Negative Cases | TN: 9,760 | FP: 140 |
| Positive Cases | FN: 40 | TP: 60 |

Now, your boss asks you three questions:

1 What percent of your predictions were correct?   
You answer: the "accuracy" was (9,760+60) out of 10,000 = 98.2%

2 What percent of the positive cases did you catch?   
You answer: the "recall" was 60 out of 100 = 60%

3 What percent of positive predictions were correct?   
You answer: the "precision" was 60 out of 200 = 30%

See also a very good explanation of [Precision and recall](https://en.wikipedia.org/wiki/Precision_and_recall) in Wikipedia.   
  
   
Fig 4: Precision and Recall.   
  
ROC curve represents a relation between sensitivity (RECALL) and specificity(NOT PRECISION) and is commonly used to measure the performance of binary classifiers. However, when dealing with highly skewed datasets, [Precision-Recall (PR) curves](http://pages.cs.wisc.edu/~jdavis/davisgoadrichcamera2.pdf) give a more representative picture of performance. See also this Quora answer: [What is the difference between a ROC curve and a precision-recall curve?](https://www.quora.com/What-is-the-difference-between-a-ROC-curve-and-a-precision-recall-curve-When-should-I-use-each). 

Q5. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything?

Answer by [Anmol Rajpurohit.](https://www.kdnuggets.com/author/anmol-rajpurohit)   
  
Often it is observed that in the pursuit of rapid innovation (aka "quick fame"), the principles of scientific methodology are violated leading to misleading innovations, i.e. appealing insights that are confirmed without rigorous validation. One such scenario is the case that given the task of improving an algorithm to yield better results, you might come with several ideas with potential for improvement.   
  
An obvious human urge is to announce these ideas ASAP and ask for their implementation. When asked for supporting data, often limited results are shared, which are very likely to be impacted by selection bias (known or unknown) or a misleading global minima (due to lack of appropriate variety in test data).   
  
Data scientists do not let their human emotions overrun their logical reasoning. While the exact approach to prove that one improvement you've brought to an algorithm is really an improvement over not doing anything would depend on the actual case at hand, there are a few common guidelines:

1 Ensure that there is no selection bias in test data used for performance comparison

2 Ensure that the test data has sufficient variety in order to be symbolic of real-life data (helps avoid overfitting)

3 Ensure that "controlled experiment" principles are followed i.e. while comparing performance, the test environment (hardware, etc.) must be exactly the same while running original algorithm and new algorithm

4 Ensure that the results are repeatable with near similar results

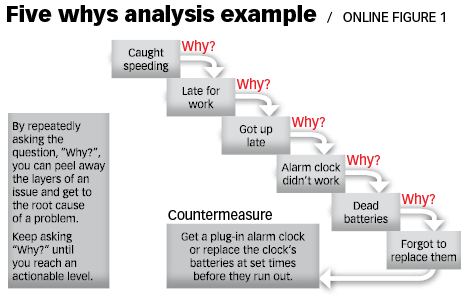
5 Examine whether the results reflect local maxima/minima or global maxima/minima

   
One common way to achieve the above guidelines is through A/B testing, where both the versions of algorithm are kept running on similar environment for a considerably long time and real-life input data is randomly split between the two. This approach is particularly common in Web Analytics. 

Q6. What is root cause analysis?

Answer by [Gregory Piatetsky](https://www.kdnuggets.com/author/gregory-piatetsky):   
  
According to Wikipedia, 

[Root cause analysis (RCA)](https://en.wikipedia.org/wiki/Root_cause_analysis) is a method of problem solving used for identifying the root causes of faults or problems. A factor is considered a root cause if removal thereof from the problem-fault-sequence prevents the final undesirable event from recurring; whereas a causal factor is one that affects an event's outcome, but is not a root cause.

Root cause analysis was initially developed to analyze industrial accidents, but is now widely used in other areas, such as healthcare, project management, or software testing.   
  
Here is a useful [Root Cause Analysis Toolkit](http://www.health.state.mn.us/patientsafety/toolkit/) from the state of Minnesota.   
  
Essentially, you can find the root cause of a problem and show the relationship of causes by repeatedly asking the question, "Why?", until you find the root of the problem. This technique is commonly called "5 Whys", although is can be involve more or less than 5 questions.   
  
   
Fig. 5 Whys Analysis Example, from [The Art of Root Cause Analysis](http://asq.org/quality-progress/2015/02/back-to-basics/the-art-of-root-cause-analysis.html) . 

Q7. Are you familiar with price optimization, price elasticity, inventory management, competitive intelligence? Give examples.

Answer by [Gregory Piatetsky](https://www.kdnuggets.com/author/gregory-piatetsky):   
  
Those are economics terms that are not frequently asked of Data Scientists but they are useful to know.   
  
[Price optimization](https://en.wikipedia.org/wiki/Price_optimization) is the use of mathematical tools to determine how customers will respond to different prices for its products and services through different channels.   
  
Big Data and data mining enables use of personalization for price optimization. Now companies like Amazon can even take optimization further and show different prices to different visitors, based on their history, although there is a strong debate about whether this is fair.   
  
Price elasticity in common usage typically refers to

1 [Price elasticity of demand](https://en.wikipedia.org/wiki/Price_elasticity_of_demand), a measure of price sensitivity. It is computed as:   
Price Elasticity of Demand = % Change in Quantity Demanded / % Change in Price.

   
Similarly, [Price elasticity of supply](https://en.wikipedia.org/wiki/Price_elasticity_of_supply) is an economics measure that shows how the quantity supplied of a good or service responds to a change in its price.   
  
[Inventory management](http://www.investopedia.com/terms/i/inventory-management.asp) is the overseeing and controlling of the ordering, storage and use of components that a company will use in the production of the items it will sell as well as the overseeing and controlling of quantities of finished products for sale.   
  
Wikipedia defines 

[Competitive intelligence](https://en.wikipedia.org/wiki/Competitive_intelligence): the action of defining, gathering, analyzing, and distributing intelligence about products, customers, competitors, and any aspect of the environment needed to support executives and managers making strategic decisions for an organization.

Tools like Google Trends, Alexa, Compete, can be used to determine general trends and analyze your competitors on the web.

8. What is statistical power?

Answer by [Gregory Piatetsky](https://www.kdnuggets.com/author/gregory-piatetsky):   
  
Wikipedia defines [Statistical power or sensitivity](https://en.wikipedia.org/wiki/Statistical_power) of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true.   
  
To put in another way, [Statistical power](http://effectsizefaq.com/2010/05/31/what-is-statistical-power/) is the likelihood that a study will detect an effect when the effect is present. The higher the statistical power, the less likely you are to make a Type II error (concluding there is no effect when, in fact, there is).   
  
Here are some tools to [calculate statistical power](https://www.dssresearch.com/KnowledgeCenter/toolkitcalculators/statisticalpowercalculators.aspx). 

9. Explain what resampling methods are and why they are useful. Also explain their limitations.

Answer by [Gregory Piatetsky](https://www.kdnuggets.com/author/gregory-piatetsky):   
  
Classical statistical parametric tests compare observed statistics to theoretical sampling distributions. Resampling a data-driven, not theory-driven methodology which is based upon repeated sampling within the same sample.   
  
Resampling refers to methods for doing one of these

1 Estimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)

2 Exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomization tests, or re-randomization tests)

3 Validating models by using random subsets (bootstrapping, cross validation)

10. Is it better to have too many false positives, or too many false negatives? Explain.

Answer by [Devendra Desale](https://www.kdnuggets.com/author/devendra-desale).   
  
It depends on the question as well as on the domain for which we are trying to solve the question.   
  
In medical testing, false negatives may provide a falsely reassuring message to patients and physicians that disease is absent, when it is actually present. This sometimes leads to inappropriate or inadequate treatment of both the patient and their disease. So, it is desired to have too many false positive.   
  
For spam filtering, a false positive occurs when spam filtering or spam blocking techniques wrongly classify a legitimate email message as spam and, as a result, interferes with its delivery. While most anti-spam tactics can block or filter a high percentage of unwanted emails, doing so without creating significant false-positive results is a much more demanding task. So, we prefer too many false negatives over many false positives. 

11. What is selection bias, why is it important and how can you avoid it?

Answer by [Matthew Mayo](https://www.kdnuggets.com/author/matt-mayo).   
  
Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample. For example, if a given sample of 100 test cases was made up of a 60/20/15/5 split of 4 classes which actually occurred in relatively equal numbers in the population, then a given model may make the false assumption that probability could be the determining predictive factor. Avoiding non-random samples is the best way to deal with bias; however, when this is impractical, techniques such as [resampling](https://en.wikipedia.org/wiki/Resampling_(statistics)), [boosting](https://en.wikipedia.org/wiki/Boosting_(machine_learning)), and weighting are strategies which can be introduced to help deal with the situation.   
  
Here is [the second part of the answers](https://www.kdnuggets.com/2016/02/21-data-science-interview-questions-answers-part2.html).

What tools or devices help you succeed in your role as a data scientist?

This question's purpose is to learn the programming languages and applications the candidate knows and has experience using. The answer will show the candidate's need for additional training of basic programming languages and platforms or any transferable skills. This is vital to understand as it can cost more time and money to train if the candidate is not knowledgeable in all of the languages and applications required for the position. Answers to look for include:

1 Experience in SAS and R programming

2 Understanding of Python, PHP or Java programming languages

3 Experience using data visualization tools

"I believe I can excel in this position with my R, Python, and SQL programming skill set. I enjoy working on the FUSE and Tableau platforms to mine data and draw inferences."

This question's purpose is to learn the programming languages and applications the candidate knows and has experience using. The answer will show the candidate's need for additional training of basic programming languages and platforms or any transferable skills. This is vital to understand as it can cost more time and money to train if the candidate is not knowledgeable in all of the languages and applications required for the position. Answers to look for include:

Experience in SAS and R programming

Understanding of Python, PHP or Java programming languages

Experience using data visualization tools

"I believe I can excel in this position with my R, Python, and SQL programming skill set. I enjoy working on the FUSE and Tableau platforms to mine data and draw inferences."

How do you identify a barrier to performance?

This question will determine how the candidate approaches solving real-world issues they will face in their role as a data scientist. It will also determine how they approach problem-solving from an analytical standpoint. This information is vital to understand because data scientists must have strong analytical and problem-solving skills. Look for answers that reveal:

1 Examples of problem-solving methods

2 Steps to take to identify the barriers to performance

3 Benchmarks for assessing performance

"My approach to determining performance bottlenecks is to conduct a performance test. I then evaluate the performance based on criteria set by the lead data scientist or company and discuss my findings with my team lead and group."

This question will determine how the candidate approaches solving real-world issues they will face in their role as a data scientist. It will also determine how they approach problem-solving from an analytical standpoint. This information is vital to understand because data scientists must have strong analytical and problem-solving skills. Look for answers that reveal:

Examples of problem-solving methods

Steps to take to identify the barriers to performance

Benchmarks for assessing performance

"My approach to determining performance bottlenecks is to conduct a performance test. I then evaluate the performance based on criteria set by the lead data scientist or company and discuss my findings with my team lead and group."

Why do you want to work at this company as a data scientist?

The purpose of this question is to determine the motivation behind the candidate's choice of applying and interviewing for the position. Their answer should reveal their inspiration for working for the company and their drive for being a data scientist. It should show the candidate is pursuing the position because they are passionate about data and believe in the company, two elements that can determine the candidate's performance. Answers to look for include:

1 Interest in data mining

2 Respect for the company's innovative practices

3 Desire to apply analytical skills to solve real-world issues with data

"I have a passion for working for data-driven, innovative companies. Your firm uses advanced technology to address everyday problems for consumers and businesses alike, which I admire. I also enjoy solving issues using an analytical approach and am passionate about incorporating technology into my work. I believe that my skills and passion match the company's drive and capabilities."

The purpose of this question is to determine the motivation behind the candidate's choice of applying and interviewing for the position. Their answer should reveal their inspiration for working for the company and their drive for being a data scientist. It should show the candidate is pursuing the position because they are passionate about data and believe in the company, two elements that can determine the candidate's performance. Answers to look for include:

Interest in data mining

Respect for the company's innovative practices

Desire to apply analytical skills to solve real-world issues with data

"I have a passion for working for data-driven, innovative companies. Your firm uses advanced technology to address everyday problems for consumers and businesses alike, which I admire. I also enjoy solving issues using an analytical approach and am passionate about incorporating technology into my work. I believe that my skills and passion match the company's drive and capabilities."

How has your prior experience prepared you for a role in data science?

This question helps determine the candidate's experience from a holistic perspective and reveals experience in demonstrating interpersonal, communication and technical skills. It is important to understand this because data scientists must be able to communicate their findings, work in a team environment and have the skills to perform the task. Here are some possible answers to look for:

1 Project management skills

2 Examples of working in a team environment

3 Ability to identify errors

A substantial response may include the following: "My experience in my previous positions has prepared me for this job by giving me the skills I need to work in a group setting, manage projects and quickly identify errors."

This question helps determine the candidate's experience from a holistic perspective and reveals experience in demonstrating interpersonal, communication and technical skills. It is important to understand this because data scientists must be able to communicate their findings, work in a team environment and have the skills to perform the task. Here are some possible answers to look for:

Project management skills

Examples of working in a team environment

Ability to identify errors

A substantial response may include the following: "My experience in my previous positions has prepared me for this job by giving me the skills I need to work in a group setting, manage projects and quickly identify errors."

How do you overcome challenges to your findings?

The reason for asking this question is to discover how well the candidate approaches solving conflicts in a team environment. Their answer shows the candidate's problem-solving and interpersonal skills in stressful situations. Understanding these skills is significant because group dynamics and business conditions change. Consider answers that:

1 Encourage discussion

2 Demonstrate leadership

3 Acknowledges recognizing and respecting different opinions

"I would acknowledge the validity of their findings. Then I would describe how I came to my conclusions using my data. I would also invite an open discussion of the results."

The reason for asking this question is to discover how well the candidate approaches solving conflicts in a team environment. Their answer shows the candidate's problem-solving and interpersonal skills in stressful situations. Understanding these skills is significant because group dynamics and business conditions change. Consider answers that:

Encourage discussion

Demonstrate leadership

Acknowledges recognizing and respecting different opinions

"I would acknowledge the validity of their findings. Then I would describe how I came to my conclusions using my data. I would also invite an open discussion of the results."

How do clean up and organize big data sets?

Data scientists frequently have to combine large amounts of information from various devices in several formats, such as data from a smartwatch or cellphone. Answers to this question will demonstrate how your candidate's methods for organizing large data. This information is important to know because data scientists need clean data to analyze information accurately to offer recommendations that solve business problems. Possible answers may include:

1 Automation tools

2 Value correction methods

3 Comprehension of data sets

"My data cleanup techniques involve determining if the data that I am collecting makes sense and correcting any values that are not logical after I have adequate information. I also use tools to help automate the cleanup process, such as Paxata."

Data scientists frequently have to combine large amounts of information from various devices in several formats, such as data from a smartwatch or cellphone. Answers to this question will demonstrate how your candidate's methods for organizing large data. This information is important to know because data scientists need clean data to analyze information accurately to offer recommendations that solve business problems. Possible answers may include:

Automation tools

Value correction methods

Comprehension of data sets

"My data cleanup techniques involve determining if the data that I am collecting makes sense and correcting any values that are not logical after I have adequate information. I also use tools to help automate the cleanup process, such as Paxata."

What methods do you use to identify outliers within a data set?

Data scientists must be able to go beyond classroom theoretical applications to real-world applications. Your candidate's answer to this question will show how they allocate their time to finding the best way to detect outliers. This information is important to know because it demonstrates the candidate's analytical skills. Look for answers that include:

1 Raw data analysis

2 Models

3 Approaches

"I like to use practical methods and analyze the raw data first. I will then think about which model will help me to detect any outliers."

Data scientists must be able to go beyond classroom theoretical applications to real-world applications. Your candidate’s answer to this question will show how they allocate their time to finding the best way to detect outliers. This information is important to know because it demonstrates the candidate’s analytical skills. Look for answers that include:

What do you mean by the term linear regression?

It is a technique in statistics to model the connection or relationship between one or multiple exploratory variables, attributed as x, and a scalar dependant variable attributed as y.

What is the difference between extrapolation and interpolation?

You have a list of values and when you want to estimate a value from two known values from the list, it’s called interpolation. When you extend known sets of facts or values to approximate a value, it’s called extrapolation.

What is the purpose of A/B testing?

The purpose of A/B testing is to generate crucial insights by testing two variables (A and B) of a purpose-driven campaign. The purpose is to identify which variable performs better than the other and achieves a set goal. This paves way for informed decisions.

How different is a mean value different from expected value?

Mean and expected values are similar but are used in different contexts. While expected values are usually referred to in a random variable context, mean values are referred in the contexts of sample population or probability distribution.

Why is it mandatory to clean a data set?

Cleaning data makes it into a format that allows data scientists to work on it. This is crucial because if data sets are not cleaned, it may lead to biased information that can alter business decisions. Over 80% of the time is spent by data scientists to clean data.

Sep 12, 2018

What are the steps involved in analytics projects?

Any analytics problem involves the following steps:

Understanding a business problem

Data exploration

Data preparation for modeling

Running the model and analysis of results

Model validation using new data sets

Model implementation and tracking of results for a set period of time

What do you understand by the term recommender systems?

Recommender systems are part of an information filtering system that is used to predict and anticipate the ratings or preferences a user is most likely to give to a product or service. You can see recommender systems at work on eCommerce websites, movie websites, research articles, music apps, news and more.

If you had to choose between the programming languages R and Python, Which one would you use for text analytics?

Personally, I would choose [Python](https://www.jigsawacademy.com/data-science-training/) for text analytics as it offers solid data analysis tools and simple data structures, thanks to its Panda library.

For linear regression, what are some of the assumptions a data scientist is most likely to make?

Some of the assumptions include the following:

Linear relationship

Multivariate normality

No auto-correlation

No or little multicollinearity

Homoscedasticity

How do you find the correlation between a categorical variable and a continuous variable?

It is possible to find the correlation between a categorical variable and a continuous variable using the analysis of covariance technique.

So, these were some of the most common analytics interview questions. Apart from these, there are also questions like coding and writing a program on languages. If you didn’t know answers to these questions, read and understand them. If you know, pass this on to people you think would benefit from this. Also, if you’ve been asked any unique analytics question, share it in the comments below.

Good luck!

# **Data Science Python Interview Questions and Answers**

The questions below are based on the course that is taught at DeZyre – [Data Science in Python](https://www.dezyre.com/data-science-in-python-online-training/36). This is not a guarantee that these questions will be asked in Data Science Interviews. The purpose of these questions is to make the reader aware of the kind of knowledge that an applicant for a Data Scientist position needs to possess.

Data Science Interview Questions in Python are generally scenario based or problem based questions where candidates are provided with a data set and asked to do data munging, data exploration, data visualization, modelling, machine learning, etc. Most of the data science interview questions are subjective and the answers to these questions vary, based on the given data problem. The main aim of the interviewer is to see how you code, what are the visualizations you can draw from the data, the conclusions you can make from the data set, etc.

**1) How can you build a simple logistic regression model in Python?**

**2) How can you train and interpret a linear regression model in SciKit learn?**

**3) Name a few libraries in Python used for Data Analysis and Scientific computations.**

NumPy, SciPy, Pandas, SciKit, Matplotlib, Seaborn

**4) Which library would you prefer for plotting in Python language: Seaborn or Matplotlib?**

Matplotlib is the python library used for plotting but it needs lot of fine-tuning to ensure that the plots look shiny. Seaborn helps data scientists create statistically and aesthetically appealing meaningful plots. The answer to this question varies based on the requirements for plotting data.

**5)  What is the main difference between a Pandas series and a single-column DataFrame in Python?**

**6) Write code to sort a DataFrame in Python in descending order.**

**7) How can you handle duplicate values in a dataset for a variable in Python?**

**8) Which Random Forest parameters can be tuned to enhance the predictive power of the model?**

**9) Which method in pandas.tools.plotting is used to create scatter plot matrix?**

    Scatter\_matrix

**10) How can you check if a data set or time series is Random?**

To check whether a dataset is random or not use the lag plot. If the lag plot for the given dataset does not show any structure then it is random.

**11) Can we create a DataFrame with multiple data types in Python? If yes, how can you do it?**

**12) Is it possible to plot histogram in Pandas without calling Matplotlib? If yes, then write the code to plot the histogram?**

**13) What are the possible ways to load an array from a text data file in Python? How can the efficiency of the code to load data file be improved?**

   numpy.loadtxt ()

**14) Which is the standard data missing marker used in Pandas?**

NaN

**15) Why you should use NumPy arrays instead of nested Python lists?**

**16)  What is the preferred method to check for an empty array in NumPy?**

**17) List down some evaluation metrics for regression problems.**

**18) Which Python library would you prefer to use for Data Munging?**

Pandas

**19) Write the code to sort an array in NumPy by the nth column?**

Using argsort () function this can be achieved. If there is an array X and you would like to sort the nth column then code for this will be x[x [: n-1].argsort ()]

**20) How are NumPy and SciPy related?**

**21) Which python library is built on top of matplotlib and Pandas to ease data plotting?**

Seaborn

**22) Which plot will you use to access the uncertainty of a statistic?**

Bootstrap

**23) What are some features of Pandas that you like or dislike?**

**24) Which scientific libraries in SciPy have you worked with in your project?**

**25) What is pylab?**

A package that combines NumPy, SciPy and Matplotlib into a single namespace.

**26) Which python library is used for Machine Learning?**

        SciKit-Learn

## **Basic Python Programming  Interview Questions**

**27) How can you copy objects in Python?**

The functions used to copy objects in Python are-

1)         Copy.copy () for shallow copy

2)         Copy.deepcopy () for deep copy

However, it is not possible to copy all objects in Python using these functions.  For instance, dictionaries have a separate copy method whereas sequences in Python have to be copied by ‘Slicing’.

**28) What is the difference between tuples and lists in Python?**

Tuples can be used as keys for dictionaries i.e. they can be hashed. Lists are mutable whereas tuples are immutable - they cannot be changed. Tuples should be used when the order of elements in a sequence matters. For example, set of actions that need to be executed in sequence, geographic locations or list of points on a specific route.

**29) What is PEP8?**

PEP8 consists of coding guidelines for Python language so that programmers can write readable code making it easy to use for any other person, later on.

**30) Is all the memory freed when Python exits?**

No it is not, because the objects that are referenced from global namespaces of Python modules are not always de-allocated when Python exits.

**31) What does \_init\_.py do?**

\_init\_.py is an empty py file used for importing a module in a directory. \_init\_.py provides an easy way to organize the files. If there is a module maindir/subdir/module.py,\_init\_.py is placed in all the directories so that the module can be imported using the following command-

import  maindir.subdir.module

**32) What is the different between range () and xrange () functions in Python?**

range () returns a list whereas xrange () returns an object that acts like an iterator for generating numbers on demand.

**33) How can you randomize the items of a list in place in Python?**

Shuffle (lst) can be used for randomizing the items of a list in Python

**34) What is a pass in Python?**

Pass in Python signifies a no operation statement indicating that nothing is to be done.

**35) If you are gives the first and last names of employees, which data type in Python will you use to store them?**

You can use a list that has first name and last name included in an element or use Dictionary.

**36) What happens when you execute the statement mango=banana in Python?**

A name error will occur when this statement is executed in Python.

**37) Write a sorting algorithm for a numerical dataset in Python.**

**38) Optimize the below python code-**

**word = 'word'**

**print word.\_\_len\_\_ ()**

**Answer:** print ‘word’.\_len\_ ()

**39) What is monkey patching in Python?**

Monkey patching is a technique that helps the programmer to modify or extend other code at runtime. Monkey patching comes handy in testing but it is not a good practice to use it in production environment as debugging the code could become difficult.

**40) Which tool in Python will you use to find bugs if any?**

Pylint and Pychecker. Pylint verifies that a module satisfies all the coding standards or not. Pychecker is a static analysis tool that helps find out bugs in the course code.

**41) How are arguments passed in Python- by reference or by value?**

The answer to this question is neither of these because passing semantics in Python are completely different. In all cases, Python passes arguments by value where all values are references to objects.

**42) You are given a list of N numbers. Create a single list comprehension in Python to create a new list that contains only those values which have even numbers from elements of the list at even indices. For instance if list[4] has an even value the it has be included in the new output list because it has an even index but if list[5] has an even value it should not be included in the list because it is not at an even index.**

 [x for x in list [1::2] if x%2 == 0]

The above code will take all the numbers present at even indices and then discard the odd numbers.

**43) Explain the usage of decorators.**

Decorators in Python are used to modify or inject code in functions or classes. Using decorators, you can wrap a class or function method call so that a piece of code can be executed before or after the execution of the original code. Decorators can be used to check for permissions, modify or track the arguments passed to a method, logging the calls to a specific method, etc.

**44) How can you check whether a pandas data frame is empty or not?**

The attribute df.empty is used to check whether a data frame is empty or not.

**45) What will be the output of the below Python code –**

**def multipliers ():**

**return [lambda x: i \* x for i in range (4)]**

**print [m (2) for m in multipliers ()]**

The output for the above code will be [6, 6,6,6]. The reason for this is that because of late binding the value of the variable i is looked up when any of the functions returned by multipliers are called.

**46) What do you mean by list comprehension?**

The process of creating a list while performing some operation on the data so that it can be accessed using an iterator is referred to as List Comprehension.

Example:

[ord (j) for j in string.ascii\_uppercase]

     [65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90]

**47)       What will be the output of the below code**

**word = ‘aeioubcdfg'**

**print word [:3] + word [3:]**

The output for the above code will be: ‘aeioubcdfg'.

In string slicing when the indices of both the slices collide and a “+” operator is applied on the string it concatenates them.

**48)       list= [‘a’,’e’,’i’,’o’,’u’]**

**print list [8:]**

The output for the above code will be an empty list []. Most of the people might confuse the answer with an index error because the code is attempting to access a member in the list whose index exceeds the total number of members in the list. The reason being the code is trying to access the slice of a list at a starting index which is greater than the number of members in the list.

**49)       What will be the output of the below code:**

**def foo (i= []):**

**i.append (1)**

**return i**

**>>> foo ()**

**>>> foo ()**

The output for the above code will be-

[1]

[1, 1]

Argument to the function foo is evaluated only once when the function is defined. However, since it is a list, on every all the list is modified by appending a 1 to it.

**50) Can the lambda forms in Python contain statements?**

No, as their syntax is restrcited to single expressions and they are used for creating function objects which are returned at runtime.

This list of questions for Python interview questions and answers is not an exhaustive one and will continue to be a work in progress. Let us know in comments below if we missed out on any important question that needs to be up here.

# Data Analyst Interview Questions to prepare for in 2018

## **Interview Questions for Data Analysts**

**1) What is the difference between Data Mining and Data Analysis?**

|  |  |
| --- | --- |
| ****Data Mining**** | ****Data Analysis**** |
| Data mining usually does not require any hypothesis. | Data analysis begins with a question or an assumption. |
| Data Mining depends on clean and well-documented data. | Data analysis involves data cleaning. |
| Results of data mining are not always easy to interpret. | Data analysts interpret the results and convey the to the stakeholders. |
| Data mining algorithms automatically develop equations. | Data analysts have to develop their own equations based on the hypothesis. |

**2) Explain the typical data analysis process.**

Data analysis deals with collecting, inspecting, cleansing, transforming and modelling data to glean valuable insights and support better decision making in an organization. The various steps involved in the data analysis process include –

**Data Exploration –**

Having identified the business problem, a data analyst has to go through the data provided by the client to analyse the root cause of the problem.

**Data Preparation**

This is the most crucial step of the data analysis process wherein any data anomalies (like missing values or detecting outliers) with the data have to be modelled in the right direction.

**Data Modelling**

The modelling step begins once the data has been prepared. Modelling is an iterative process wherein the model is run repeatedly for improvements. Data modelling ensures that the best possible result is found for a given business problem.

**Validation**

In this step, the model provided by the client and the model developed by the data analyst are validated against each other to find out if the developed model will meet the business requirements.

**Implementation of the Model and Tracking**

This is the final step of the data analysis process wherein the model is implemented in production and is tested for accuracy and efficiency.

**3) What is the difference between Data Mining and Data Profiling?**

Data Profiling, also referred to as Data Archeology is the process of assessing the data values in a given dataset for uniqueness, consistency and logic. Data profiling cannot identify any incorrect or inaccurate data but can detect only business rules violations or anomalies. The main purpose of data profiling is to find out if the existing data can be used for various other purposes.

Data Mining refers to the analysis of datasets to find relationships that have not been discovered earlier. It focusses on sequenced discoveries or identifying dependencies, bulk analysis, finding various types of attributes, etc.

**4) How often should you retrain a data model?**

A good data analyst is the one who understands how changing business dynamics will affect the efficiency of a predictive model. You must be a valuable consultant who can use analytical skills and business acumen to find the root cause of business problems.

The best way to answer this question would be to say that you would work with the client to define a time period in advance. However, I would refresh or retrain a model when the company enters a new market, consummate an acquisition or is facing emerging competition. As a data analyst, I would retrain the model as quick as possible to adjust with the changing behaviour of customers or change in market conditions.

**5) What is data cleansing? Mention few best practices that you have followed while data cleansing.**

From a given dataset for analysis, it is extremely important to sort the information required for data analysis. Data cleaning is a crucial step in the analysis process wherein data is inspected to find any anomalies, remove repetitive data, eliminate any incorrect information, etc. Data cleansing does not involve deleting any existing information from the database, it just enhances the quality of data so that it can be used for analysis.

Some of the best practices for data cleansing include –

* i. Developing a data quality plan to identify where maximum data quality errors occur so that you can assess the root cause and design the plan according to that.
* Ii. Follow a standard process of verifying the important data before it is entered into the database.
* Iii. Identify any duplicates and validate the accuracy of the data as this will save lot of time during analysis.
* Iv. Tracking all the cleaning operations performed on the data is very important so that you repeat or remove any operations as necessary.

**6) How will you handle the QA process when developing a predictive model to forecast customer churn?**

Data analysts require inputs from the business owners and a collaborative environment to operationalize analytics. To create and deploy predictive models in production there should be an effective, efficient and repeatable process. Without taking feedback from the business owner, the model will just be a one-and-done model.

The best way to answer this question would be to say that you would first partition the data into 3 different sets Training, Testing and Validation. You would then show the results of the validation set to the business owner by eliminating biases from the first 2 sets. The input from the business owner or the client will give you an idea on whether you model predicts customer churn with accuracy and provides desired results.

**7) Mention some common problems that data analysts encounter during analysis.**

* i. Having a poor formatted data file. For instance, having CSV data with un-escaped newlines and commas in columns.
* Ii. Having inconsistent and incomplete data can be frustrating.
* Iii. Common Misspelling and Duplicate entries are a common data quality problem that most of the data analysts face.
* Iv. Having different value representations and misclassified data.

**8) What are the important steps in data validation process?**

Data Validation is performed in 2 different steps-

Data Screening – In this step various algorithms are used to screen the entire data to find any erroneous or questionable values. Such values need to be examined and should be handled.

Data Verification- In this step each suspect value is evaluated on case by case basis and a decision is to be made if the values have to be accepted as valid or if the values have to be rejected as invalid or if they have to be replaced with some redundant values.

**9) How will you create a classification to identify key customer trends in unstructured data?**

A model does not hold any value if it cannot produce actionable results, an experienced data analyst will have a varying strategy based on the type of data being analysed. For example, if a customer complain was retweeted then should that data be included or not. Also, any sensitive data of the customer needs to be protected, so it is also advisable to consult with the stakeholder to ensure that you are following all the compliance regulations of the organization and disclosure laws, if any.

You can answer this question by stating that you would first consult with the stakeholder of the business to understand the objective of classifying this data. Then, you would use an iterative process by pulling new data samples and modifying the model accordingly and evaluating it for accuracy. You can mention that you would follow a basic process of mapping the data, creating an algorithm, mining the data, visualizing it and so on. However, you would accomplish this in multiple segments by considering the feedback from stakeholders to ensure that you develop an enriching model that can produce actionable results.

**10) What is the criteria to say whether a developed data model is good or not?**

* i.The developed model should have predictable performance.
* Ii.A good data model can adapt easily to any changes in business requirements.
* Iii.Any major data changes in a good data model should be scalable.
* Iv.A good data model is one that can be easily consumed for actionable results.

**11) According to you what are the qualities/skills that a data analyst must posses to be successful at this position.**

Problem Solving and Analytical thinking are the two important skills to be successful as a data analyst. One needs to skilled ar formatting data so that the gleaned information is available in a easy-to-read manner. Not to forget technical proficiency is of significant importance. You can also talk about other skills that the interviewer expects in an ideal candidate for the job position based on the given job description.

### ****12) You are assigned a new data anlytics project. How will you begin with and what are the steps you will follow?****

The purpose of asking this question is that the interviewer wants to understand how you approach a given data problem and what is the though process you follow to ensure that you are organized. You can start answering this question by saying that you will start with finding the objective of the given problem and defining it so that there is solid direction on what need to be done. The next step would be to do data exploration and familiarise myself with the entire dataset which is very important when working with a new dataset.The next step would be to prepare the data for modelling which would including finding outliers, handling missing values and validating the data. Having validated the data, I will start data modelling untill I discover any meaningfuk insights. After this the final step would be to implement the model and track the output results.

This is the generic data analysis process that we have explained in this answer, however, the answer to your  question might slightly change based on the kind of data problem and the tools available at hand.

**13) What do you know about  interquartile range as data analyst?**

A measure of the dispersion of data that is shown in a box plot is referred to as the interquartile range. It is the difference between the upper and the lower quartile.

Data Science Interviews at Top Tech Companies

Facebook Data Science Interview Questions

1)         A building has 100 floors. Given 2 identical eggs, how can you use them to find the threshold floor? The egg will break from any particular floor above floor N, including floor N itself.

2)         In a given day, how many birthday posts occur on Facebook?

3)         You are at a Casino. You have two dices to play with. You win $10 every time you roll a 5. If you play till you win and then stop, what is the expected pay-out?

4)         How many big Macs does McDonald sell every year in US?

5)         You are about to get on a plane to Seattle, you want to know whether you have to bring an umbrella or not. You call three of your random friends and as each one of them if it’s raining. The probability that your friend is telling the truth is 2/3 and the probability that they are playing a prank on you by lying is 1/3. If all 3 of them tell that it is raining, then what is the probability that it is actually raining in Seattle.

6)         You can roll a dice three times. You will be given $X where X is the highest roll you get. You can choose to stop rolling at any time (example, if you roll a 6 on the first roll, you can stop). What is your expected pay-out?

7)         How can bogus Facebook accounts be detected?

8)       You have been given the data on Facebook user’s friending or defriending each other. How will you determine whether a given pair of Facebook users are friends or not?

9)         How many dentists are there in US?

10)         You have 2 dices. What is the probability of getting at least one 4? Also find out the probability of getting at least one 4 if you have n dices.

11)       Pick up a coin C1 given C1+C2 with probability of trials p (h1) =.7, p (h2) =.6 and doing 10 trials. And what is the probability that the given coin you picked is C1 given you have 7 heads and 3 tails?

12)     You are given two tables- friend\_request and request\_accepted. Friend\_request contains requester\_id, time and sent\_to\_id and request\_accepted table contains time, acceptor\_id and requestor\_id. How will you determine the overall acceptance rate of requests?

13)       How would add new Facebook members to the database of members, and code their relationships to others in the database?

14)       What would you add to Facebook and how would you pitch it and measure its success?

15)  How will you test that there is increased probability of a user to stay active after 6 months given that a user has more friends now?

16) You have two tables-the first table has data about the users and their friends, the second table has data about the users and the pages they have liked. Write an SQL query to make recommendations using pages that your friends liked. The query result should not recommend the pages that have already been liked by a user.

17) What is the probability of pulling a different shape or a different colour card from a deck of 52 cards?

18) Which technique will you use to compare the performance of two back-end engines that generate automatic friend recommendations on Facebook?

19) Implement a sorting algorithm for a numerical dataset in Python.

20) How many people are using Facebook in California at 1.30 PM on Monday?

21) You are given 50 cards with five different colors- 10 Green cards, 10 Red Cards, 10 Orange Cards, 10 Blue cards, and 10 Yellow cards. The cards of each colors are numbered from one to ten. Two cards are picked at random. Find out the probability that the cards picked are not of same number and same color.

22) What approach will you follow to develop the love,like, sad feature on Facebook?

Insight Data Science Interview Questions

1)         Which companies participating in Insight would you be interested in working for?

2)         Create a program in a language of your choice to read a text file with various tweets. The output should be 2 text files-one that contains the list of all unique words among all tweets along with the count for repeated words and the second file should contain the medium number of unique words for all tweets.

3)         What motivates you to transition from academia to data science?

Twitter Data Scientist Interview Questions

1)    How can you measure engagement with given Twitter data?

2)    Give a large dataset, find the median.

3)    What is the good measure of influence of a Twitter user?

AirBnB Data Science Interview Questions

1)  Do you have some knowledge of R - analyse a given dataset in R?

2)  What will you do if removing missing values from a dataset cause bias?

3)  How can you reduce bias in a given data set?

4) How will you impute missing information in a dataset?

Google Data Science Interview Questions

1)  Explain about string parsing in R language

2) A disc is spinning on a spindle and you don’t know the direction in which way the disc is spinning. You are provided with a set of pins.How will you use the pins to describe in which way the disc is spinning?

3)  Describe the data analysis process.

4) How will you cut a circular cake into 8 equal pieces?

LinkedIn Data Science Interview Questions

1)  Find out K most frequent numbers from a given stream of numbers on the fly.

2)  Given 2 vectors, how will you generate a sorted vector?

3)  Implementing pow function

4)  What kind of product you want to build at LinkedIn?

5)  How will you design a recommendation engine for jobs?

6)  Write a program to segment a long string into a group of valid words using Dictionary. The result should return false if the string cannot be segmented. Also explain about the complexity of the devised solution.

7) Define an algorithm to discover when a person is starting to search for new job.

8) What are the factors used to produce “People You May Know” data product on LinkedIn?

9)  How will you find the second largest element in a Binary Search tree ? (Asked for a Data Scientist Intern job role)

Mu Sigma Data Science Interview Questions

1)   Explain the difference between Supervised and Unsupervised Learning through examples.

2)   How would you add value to the company through your projects?

3)   Case Study based questions – Cars are implanted with speed tracker so that the insurance companies can track about our driving state. Based on this new scheme what kind of business questions can be answered?

4)  Define standard deviation, mean, mode and median.

5) What is a joke that people say about you and how would you rate the joke on a scale of 1 to 10?

6) You own a clothing enterprise and want to improve your place in the market. How will you do it from the ground level ?

7) How will you customize the menu for Cafe Coffee Day ?

Amazon Data Science Interview Questions

1) Estimate the probability of a disease in a particular city given that the probability of the disease on a national level is low.

2) How will inspect missing data and when are they important for your analysis?

3) How will you decide whether a customer will buy a product today or not given the income of the customer, location where the customer lives, profession and gender? Define a machine learning algorithm for this.

4) From a long sorted list and a short 4 element sorted list, which algorithm will you use to search the long sorted list for 4 elements.

5) How can you compare a neural network that has one layer, one input and output to a logistic regression model?

6) How do you treat colinearity?

7) How will you deal with unbalanced data where the ratio of negative and positive is huge?

8) What is the difference between -

i) Stack and Queue

ii) Linkedin and Array

Uber Data Science Interview Questions

1) Will Uber cause city congestion?

2) What are the metrics you will use to track if Uber’s paid advertising strategies to acquire customers work? How will you figure out the acceptable cost of customer acquisition?

3) Explain principal components analysis with equations.

4) Explain about the various time series forecasting technqiues.

5) Which machine learning algorithm will you use to solve a Uber driver accepting  request?

6)How will you compare the results of various machine learning algorithms?

7) How to solve multi-collinearity?

8) How will you design the heatmap for Uber drivers to provide recommendation on where to wait for passengers? How would you approach this?

9) If we added one rider to the current SF market, how would that affect the existing riders and drivers?

10) What are the different performance metrics for evaluating Uber services?

11) How will you decide which version (Version 1 or Version 2) of the Surge Pricing Algorithms is working better for Uber ?

12) How will you explain JOIN function in SQL to a 10 year old ?

Netflix Data Science Interview Questions

1) How can you build and test a metric to compare ranked list of TV shows or Movies for two Netflix users?

2) How can you decide if one algorithm is better than the other?

Microsoft Data Science Interview Questions

1) Write a function to check whether a particular word is a palindrome or not.

2) How can you compute an inverse matrix faster by playing with some computation tricks?

3) You have a bag with 6 marbles. One marble is white.  You reach the bag 100 times. After taking out a marble, it is placed back in the bag. What is the probability of drawing a white marble at least once?

Apple Data Science Interview Questions

1) How do you take millions of users with 100's of transactions each, amongst 10000's of products and group the users together in a meaningful segments?

Adobe Data Scientist Interview Questions

1) Check whether a given integer is a palindrome or not without converting it to a string.

2) What is the degree of freedom for lasso?

3) You have two sorted array of integers, write a program to find a number from each array such that the sum of the two numbers is closest to an integer i.

American Express Data Scientist Interview Questions

1) Suppose that American Express has 1 million card members along with their transaction details. They also have 10,000 restaurants and 1000 food coupons. Suggest a method which can be used to pass the food coupons to users given that some users have already received the food coupons so far.

2) You are given a training dataset of users that contain their demographic details, the pages on Facebook they have liked so far and results of psychology test  based on their personality i.e. their openness to like FB pages or not. How will you predict the age, gender and other demographics of unseen data?

Quora Data Scientist Interview Questions

1) How will you test a machine learning model for accuracy?

2) Print the elements of a matrix in zig-zag manner.

3) How will you overcome overfitting in predictive models?

4) Develop an algorithm to sort two lists of sorted integers into a single list.

Goldman Sachs Data Scientist Interview Questions

1) Count the total number of trees in United States.

2) Estimate the number of square feet pizza’s eaten in US each year.

3) A box has 12 red cards and 12 black cards. Another box has 24 red cards and 24 black cards. You want to draw two cards at random from one of the two boxes, which box has a higher probability of getting cards of same colour and why?

4) How will you prove that the square root of 2 is irrational?

5) What is the probability of getting a HTT combination before getting a TTH combination?

6) There are 8 identical balls and only one of the ball is slightly heavier than the others. You are given a balance scale to find the heavier ball. What is the least number of times you have to use the balance scale to find the heavier ball?

Walmart Data Science Interview Questions

1) Write the code to reverse a Linked list.

2) What assumptions does linear regression machine learning algorithm make?

3) A stranger uses a search engine to find something and you do not know anything about the person. How will you design an algorithm to determine what the stranger is looking for just after he/she types few characters in the search box?

4) How will you fix multi-colinearity in a regression model?

5) What data structures are available in the Pandas package in Python programming language?

6) State some use cases where Hadoop MapReduce works well and where it does not.

7) What is the difference between an iterator, generator and list comprehension in Python?

8) What is the difference between a bagged model and a boosted model?

9) What do you understand by parametric and non-parametric methods? Explain with examples.

10) Have you used sampling? What are the various types of sampling have you worked with?

11) Explain about cross entropy ?

12) What are the assuptions you make for linear regression ?

13) Differentiate between gradient boosting and random forest.

14) What is the signigicance of log odds ?

IBM Data Science Interview Questions

1) How will you handle missing data ?

Yammer Data Science Interview Questions

1 How can you solve a problem that has no solution?

2 On rolling a dice if you get $1 per dot on the upturned face,what are your expected earnings from rolling a dice?

3 In continuation with question #2, if you have 2 chances to roll the dice and you are given the opportunity to decide when to stop rolling the dice (in the first roll or in the second roll). What will be your rolling strategy to get maximum earnings?

4 What will be your expected earnings with the two roll strategy?

5 You are creating a report for user content uploads every month and observe a sudden increase in the number of upload for the month of November. The increase in uploads is particularly in image uploads. What do you think will be the cause for this and how will you test this sudden spike?

Citi Bank Data Science Interview Questions

1) A dice is rolled twice, what is the probability that on the second chance it will be a 6?

2) What are Type 1 and Type 2 errors ?

3) Burn two ropes, one needs 60 minutes of time to burn and the other needs 30 minutes of time. How will you achieve this in 45 minutes of time ?

Data Science Interview Questions Asked at Other Top Tech Companies

1) R programming language cannot handle large amounts of data. What are the other ways of handling it without using Hadoop infrastructure? (Asked at Pyro Networks)

2) Explain the working of a Random Forest Machine Learning Algorithm (Asked at Cyient)

3) Describe K-Means Clustering.(Asked at Symphony Teleca)

4) What is the difference between logistic and linear regression? (Asked at Symphony Teleca)

5) What kind of distribution does logistic regression follow? (Asked at Symphony Teleca)

6) How do you parallelize machine learning algorithms? (Asked at Vodafone)

7) When required data is not available for analysis, how do you go about collecting it? (Asked at Vodafone)

8) What do you understand by heteroscadisticity (Asked at Vodafone)

9) What do you understand by confidence interval? (Asked at Vodafone)

10) Difference between adjusted r and r square. (Asked at Vodafone)

11) How Facebook recommends items to newsfeed? (Asked at Finomena)

12)  What do you understand by ROC curve and how is it used? (Asked at MachinePulse)

13) How will you identify the top K queries from a file? (Asked at BloomReach)

14) Given a set of webpages and changes on the website, how will you test the new website feature to determine if the change works positively? (Asked at BloomReach)

15) There are N pieces of rope in a bucket. You put your hand into the bucket, take one end piece of the rope .Again you put your hand into the bucket and take another end piece of a rope. You tie both the end pieces together. What is the expected value of the number of loops within the bucket? (Asked at Natera)

16) How will you test if a chosen credit scoring model works or not? What data will you look at? (Asked at Square)

17) There are 10 bottles where each contains coins of 1 gram each. There is one bottle of that contains 1.1 gram coins. How will you identify that bottle after only one measurement? (Data Science Puzzle asked at Latent View Analytics)

18) How will you measure a cylindrical glass filled with water whether it is exactly half filled or not? You cannot measure the water, you cannot measure the height of the glass nor can you dip anything into the glass. (Data Science Puzzle asked at Latent View Analytics)

19) What would you do if you were a traffic sign? (Data Science Interview Question asked at Latent View Analytics)

20)  If you could get the dataset on any topic of interest, irespective of the collection methods or resources then how would the dataset look like and what will you do with it. (Data Scientist Interview Question asked at CKM Advisors)

21) Given n samples from a uniform distribution [0,d], how will you estimate the value of d? (Data Scientist Interview Question asked at Spotify)

22) How will you tune a Random Forest? (Data Science Interview Question asked at Instacart).

23) Tell us about a project where you have extracted useful information from a large dataset. Which machine learning algorithm did you use for this and why? (Data Scientist Interview Question asked at Greenplum)

24) What is the difference between Z test and T test ? (Data Scientist Interview Questions asked at Antuit)

25) What are the different models you have used for analysis and what were your inferences? (Data Scientist Interview Questions asked at Cognizant)

26) Given the title of a product, identify the category and sub-category of the product. (Data Scientist interview question asked at Delhivery)

27) What is the difference between machine learning and deep learning? ( Data Scientist Interview Question asked at InfoObjects)

28) What are the different parameters in ARIMA models ? (Data Science Interview Question asked at Morgan Stanley)

29) What are the optimisations you would consider when computing the similarity matrix for a large dataset? (Data Science Interview questions asked at MakeMyTrip)

30) Use Python programming language to implement a toolbox with specific image processing tasks.(Data Science Interview Question asked at Intuitive Surgical)

31) Why do you use Random Forest instead of a simple classifier for one of the classification problems ? (Data Science Interview Question asked at Audi)

32) What is an n-gram? (Data Science Interview Question asked at Yelp)

33) What are the problems related to Overfitting and Underfitting  and how will you deal with these ? (Data Science Interview Question asked at Tiger Analytics)

34) Given a MxN dimension matrix with each cell containing an alphabet, find if a string is contained in it or not.(Data Science Interview Question asked at Tiger Analytics)

35) How do you "Group By" in R programming language without making use of any package ? (Data Scientist Interview Question asked at OLX)

36) List 15 features that you will make use of to build a classifier for OLX website.(Data Scientist Interview Question asked at OLX)

37) How will you build a caching system using an advanced data structure like hashmap ? (Data Scientist Interview Question asked at OLX)

38) How to reverse strings that have changing positions ? (Data Scientist Interview Question asked at Tiger Analytics)

39) How do you select a cricket team ? (Data Scientist Interview Question asked at Quantiphi)

40) What is the difference between trees and random forest ? (Data Scientist Interview Question asked at Salesforce)

If you are asked questions like what is your favourite leisure activity? Or something like what is that you like to do for fun?  Most of the people often tend to answer that they like to read programming books or do coding thinking that this is what they are supposed to say in a technical interview. Is this something you really do it for fun? A key point to bear in mind that the interviewer is also a person and interact with them as a person naturally. This will help the interviewer see you as an all-rounder who can visualize the company’s whole vision and not just view business problems from an academic viewpoint.

Raw data analysis

Models

Approaches

"I like to use practical methods and analyze the raw data first. I will then think about which model will help me to detect any outliers."

Questions & Solution

1) Which of these measures are used to analyze the central tendency of data?

A) Mean and Normal Distribution

B) Mean, Median and Mode

C) Mode, Alpha & Range

D) Standard Deviation, Range and Mean

E) Median, Range and Normal Distribution

Solution: (B)

The mean, median, mode are the three statistical measures which help us to analyze the central tendency of data. We use these measures to find the central value of the data to summarize the entire data set.

2) Five numbers are given: (5, 10, 15, 5, 15). Now, what would be the sum of deviations of individual data points from their mean?

A) 10

B)25

C) 50

D) 0

E) None of the above

Solution: (D)

The sum of deviations of the individual will always be 0.

3) A test is administered annually. The test has a mean score of 150 and a standard deviation of 20. If Ravi’s z-score is 1.50, what was his score on the test?

A) 180  
B) 130  
C) 30  
D) 150  
E) None of the above

Solution: (A)

X= μ+Zσ where μ is the mean,  σ is the standard deviation and X is the score we’re calculating. Therefore X = 150+20\*1.5 = 180

4) Which of the following measures of central tendency will always change if a single value in the data changes?

A) Mean

B) Median

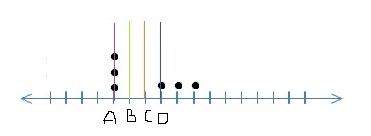
C) Mode

D) All of these

Solution: (A)

The mean of the dataset would always change if we change any value of the data set. Since we are summing up all the values together to get it, every value of the data set contributes to its value. Median and mode may or may not change with altering a single value in the dataset.

5) Below, we have represented six data points on a scale where vertical lines on scale represent unit.

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/02/27134903/Image_21.jpg)

Which of the following line represents the mean of the given data points, where the scale is divided into same units?

A) A  
B) B  
C) C  
D) D

Solution: (C)

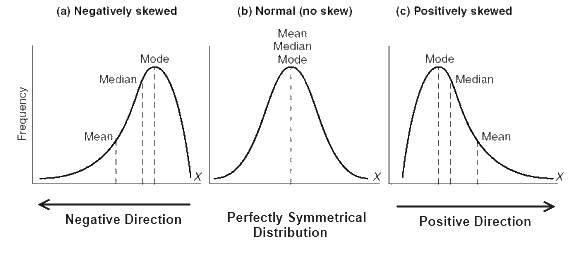
It’s a little tricky to visualize this one by just looking at the data points. We can simply substitute values to understand the mean. Let A be 1, B be 2, C be 3 and so on. The data values as shown will become {1,1,1,4,5,6} which will have mean to be 18/6 = 3 i.e. C.

6) If a positively skewed distribution has a median of 50, which of the following statement is true?

A) Mean is greater than 50  
B) Mean is less than 50  
C) Mode is less than 50  
D) Mode is greater than 50  
E) Both A and C  
F) Both B and D

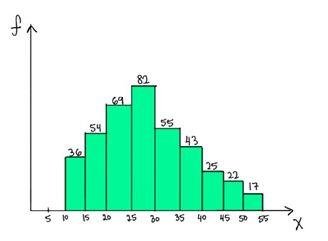
Solution: (E)

Below are the distributions for Negatively, Positively and no skewed curves.



As we can see for a positively skewed curve, Mode<Median<Mean. So if median is 50, mean would be more than 50 and mode will be less than 50.

7) Which of the following is a possible value for the median of the below distribution?

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/02/27135507/Image_41.jpg)

A) 32  
B) 26  
C) 17  
D) 40

Solution: (B)

To answer this one we need to go to the basic definition of a median. Median is the value which has roughly half the values before it and half the values after. The number of values less than 25 are (36+54+69 = 159) and the number of values greater than 30 are (55+43+25+22+17= 162). So the median should lie somewhere between 25 and 30. Hence 26 is a possible value of the median.

8) Which of the following statements are true about Bessels Correction while calculating a sample standard deviation?

Bessels correction is always done when we perform any operation on a sample data.

Bessels correction is used when we are trying to estimate population standard deviation from the sample.

Bessels corrected standard deviation is less biased.

A)  Only 2

B) Only 3

C) Both 2 and 3

D) Both 1 and 3

Solution: (C)

Contrary to the popular belief Bessel’s correction should not be always done. It’s basically done when we’re trying to estimate the population standard deviation using the sample standard deviation. The bias is definitely reduced as the standard deviation will now(after correction) be depicting the dispersion of the population more than that of the sample.

9) If the variance of a dataset is correctly computed with the formula using (n – 1) in the denominator, which of the following option is true?

A) Dataset is a sample  
B) Dataset is a population  
C) Dataset could be either a sample or a population  
D) Dataset is from a census  
E) None of the above

Solution: (A)

If the variance has n-1 in the formula, it means that the set is a sample. We try to estimate the population variance by dividing the sum of squared difference with the mean with n-1.

When we have the actual population data we can directly divide the sum of squared differences with n instead of n-1.

10) [True or False] Standard deviation can be negative.

A) TRUE

B) FALSE

Solution: (B)

Below is the formula for standard deviation

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/05/01140509/sd.png

Since the differences are squared, added and then rooted, negative standard deviations are not possible.

11) Standard deviation is robust to outliers?

A) True

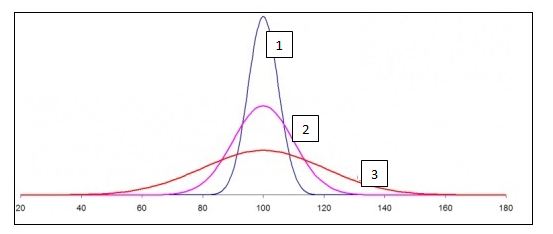
B) False

Solution: (B)

If you look at the formula for standard deviation above, a very high or a very low value would increase standard deviation as it would be very different from the mean. Hence outliers will effect standard deviation.

12) For the below normal distribution, which of the following option holds true ?

σ1, σ2 and σ3 represent the standard deviations for curves 1, 2 and 3 respectively.

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/02/27140006/Image_8.jpg)  
A) σ1> σ2> σ3

B) σ1< σ2< σ3

C) σ1= σ2= σ3

D) None

Solution: (B)

From the definition of normal distribution, we know that the area under the curve is 1 for all the 3 shapes. The curve 3 is more spread and hence more dispersed (most of values being within 40-160). Therefore it will have the highest standard deviation. Similarly, Curve 1 has a very low range and all the values are in a small range of 80-120. Hence, curve 1 has the least standard deviation.

13) What would be the critical values of Z for 98% confidence interval for a two-tailed test ?

A) +/- 2.33  
B) +/- 1.96  
C) +/- 1.64  
D) +/- 2.55

Solution: (A)

We need to look at the z table for answering this. For a 2 tailed test, and a 98% confidence interval, we should check the area before the z value as 0.99 since 1% will be on the left side of the mean and 1% on the right side. Hence we should check for the z value for area>0.99. The value will be +/- 2.33

14) [True or False] The standard normal curve is symmetric about 0 and the total area under it is 1.

A)TRUE

B) FALSE

Solution: (A)

By the definition of the normal curve, the area under it is 1 and is symmetric about zero. The mean, median and mode are all equal and 0. The area to the left of mean is equal to the area on the right of mean. Hence it is symmetric.

Context for Questions 15-17

Studies show that listening to music while studying can improve your memory. To demonstrate this, a researcher obtains a sample of 36 college students and gives them a standard memory test while they listen to some background music. Under normal circumstances (without music), the mean score obtained was 25 and standard deviation is 6. The mean score for the sample after the experiment (i.e With music) is 28.

15) What is the null hypothesis in this case?

A) Listening to music while studying will not impact memory.  
B) Listening to music while studying may worsen memory.  
C) Listening to music while studying may improve memory.  
D) Listening to music while studying will not improve memory but can make it worse.

Solution: (D)

The null hypothesis is generally assumed statement, that there is no relationship in the measured phenomena. Here the null hypothesis would be that there is no relationship between listening to music and improvement in memory.

16) What would be the Type I error?

A) Concluding that listening to music while studying improves memory, and it’s right.  
B) Concluding that listening to music while studying improves memory when it actually doesn’t.  
C) Concluding that listening to music while studying does not improve memory but it does.

Solution: (B)

Type 1 error means that we reject the null hypothesis when its actually true. Here the null hypothesis is that music does not improve memory. Type 1 error would be that we reject it and say that music does improve memory when it actually doesn’t.

17) After performing the Z-test, what can we conclude \_\_\_\_ ?

A) Listening to music does not improve memory.

B)Listening to music significantly improves memory at p

C) The information is insufficient for any conclusion.

D) None of the above

Solution: (B)

Let’s perform the Z test on the given case. We know that the null hypothesis is that listening to music does not improve memory.

Alternate hypothesis is that listening to music does improve memory.

In this case the standard error i.e. https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/05/02063602/q171.png

The Z score for a sample mean of 28 from this population is

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/05/02063716/q18-300x42.png

Z critical value for α = 0.05 (one tailed) would be 1.65 as seen from the z table.

Therefore since the Z value observed is greater than the Z critical value, we can reject the null hypothesis and say that listening to music does improve the memory with 95% confidence.

18) A researcher concludes from his analysis that a placebo cures AIDS. What type of error is he making?

A) Type 1 error

B) Type 2 error

C) None of these. The researcher is not making an error.

D) Cannot be determined

Solution: (D)

By definition, type 1 error is rejecting the null hypothesis when its actually true and type 2 error is accepting the null hypothesis when its actually false. In this case to define the error, we need to first define the null and alternate hypothesis.

19) What happens to the confidence interval when we introduce some outliers to the data?

A) Confidence interval is robust to outliers

B) Confidence interval will increase with the introduction of outliers.

C) Confidence interval will decrease with the introduction of outliers.

D) We cannot determine the confidence interval in this case.

Solution: (B)

We know that confidence interval depends on the standard deviation of the data. If we introduce outliers into the data, the standard deviation increases, and hence the confidence interval also increases.

Context for questions 20- 22

A medical doctor wants to reduce blood sugar level of all his patients by altering their diet. He finds that the mean sugar level of all patients is 180 with a standard deviation of 18. Nine of his patients start dieting and the mean of the sample is observed to 175. Now, he is considering to recommend all his patients to go on a diet.

Note: He calculates 99% confidence interval.

20) What is the standard error of the mean?

A) 9  
B) 6  
C) 7.5  
D) 18

Solution: (B)

The standard error of the mean is the standard deviation by the square root of the number of values. i.e.

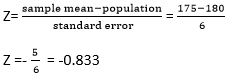
Standard error =  https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/05/02064935/q204.png = 6

21) What is the probability of getting a mean of 175 or less after all the patients start dieting?

A) 20%  
B) 25%  
C) 15%  
D) 12%

Solution: (A)

This actually wants us to calculate the probability of population mean being 175 after the intervention. We can calculate the Z value for the given mean.



If we look at the z table, the corresponding value for z = -0.833 ~ 0.2033.

Therefore there is around 20% probability that if everyone starts dieting, the population mean would be 175.

22) Which of the following statement is correct?

A) The doctor has a valid evidence that dieting reduces blood sugar level.

B) The doctor does not have enough evidence that dieting reduces blood sugar level.

C) If the doctor makes all future patients diet in a similar way, the mean blood pressure will fall below 160.

Solution: (B)

We need to check if we have sufficient evidence to reject the null. The null hypothesis is that dieting has no effect on blood sugar. This is a two tailed test. The z critical value for a 2 tailed test would be ±2.58.

The z value as we have calculated is -0.833.

Since Z value < Z critical value, we do not have enough evidence that dieting reduces blood sugar.

Question Context 23-25

A researcher is trying to examine the effects of two different teaching methods. He divides 20 students into two groups of 10 each. For group 1, the teaching method is using fun examples. Where as for group 2 the teaching method is using software to help students learn. After a 20 minutes lecture of both groups, a test is conducted for all the students.

We want to calculate if there is a significant difference in the scores of both the groups.

It is given that:

Alpha=0.05, two tailed.

Mean test score for group 1 = 10

Mean test score for group 2 = 7

Standard error = 0.94

23) What is the value of t-statistic?

A) 3.191  
B) 3.395  
C) Cannot be determined.  
D) None of the above

Solution: (A)

The t statistic of the given group is nothing but the difference between the group means by the standard error.

=(10-7)/0.94 = 3.191

24) Is there a significant difference in the scores of the two groups?

A) Yes  
B) No

Solution: (A)

The null hypothesis in this case would be that there is no difference between the groups, while the alternate hypothesis would be that the groups are significantly different.

The t critical value for a 2 tailed test at α = 0.05 is ±2.101. The t statistic obtained is 3.191. Since the t statistic is more than the critical value of t, we can reject the null hypothesis and say that the two groups are significantly different with 95% confidence.

25) What percentage of variability in scores is explained by the method of teaching?

A) 36.13  
B) 45.21  
C) 40.33  
D) 32.97

Solution: (A)

The % variability in scores is given by the R2 value. The formula for R2 given by

R2 =  https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/05/02065514/q23.png

The degrees of freedom in this case would be 10+10 -2 since there are two groups with size 10 each. The degree of freedom is 18.

R2 =  https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/05/02071945/q31.png   = 36.13

26) [True or False] F statistic cannot be negative.

A) TRUE

B) FALSE

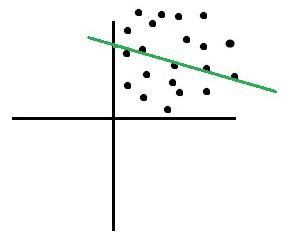
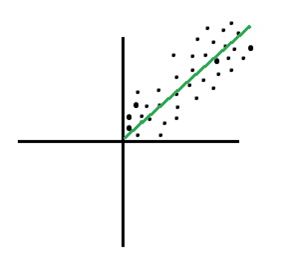
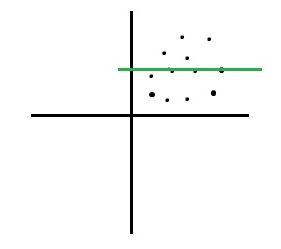
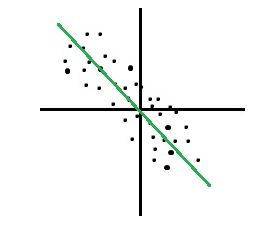
Solution: (A)

F statistic is the value we receive when we run an ANOVA test on different groups to understand the differences between them. The F statistic is given by the ratio of between group variability to within group variability

Below is the formula for f Statistic.

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/05/02072328/q301.png

Since both the numerator and denominator possess square terms, F statistic cannot be negative.

27) Which of the graph below has very strong positive correlation?  
A)[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/03/01060922/Image_A.jpg)  
B)[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/03/01061036/Image_b.jpg)  
C)[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/03/01061135/Image_c.jpg)  
D)[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/03/01061227/Image_D.jpg)

Solution: (B)

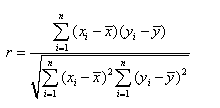
A strong positive correlation would occur when the following condition is met. If x increases, y should also increase, if x decreases, y should also decrease. The slope of the line would be positive in this case and the data points will show a clear linear relationship. Option B shows a strong positive relationship.

28) Correlation between two variables (Var1 and Var2) is 0.65. Now, after adding numeric 2 to all the values of Var1, the correlation co-efficient will\_\_\_\_\_\_\_ ?

A) Increase  
B) Decrease  
C) None of the above

Solution: (C)

If a constant value is added or subtracted to either variable, the correlation coefficient would be unchanged. It is easy to understand if we look at the formula for calculating the correlation.



If we add a constant value to all the values of x, the xi and  will change by the same number, and the differences will remain the same. Hence, there is no change in the correlation coefficient.

29) It is observed that there is a very high correlation between math test scores and amount of physical exercise done by a student on the test day. What can you infer from this?

High correlation implies that after exercise the test scores are high.

Correlation does not imply causation.

Correlation measures the strength of linear relationship between amount of exercise and test scores.

A) Only 1  
B) 1 and 3  
C) 2 and 3  
D) All the statements are true

Solution: (C)

Though sometimes causation might be intuitive from a high correlation but actually correlation does not imply any causal inference. It just tells us the strength of the relationship between the two variables. If both the variables move together, there is a high correlation among them.

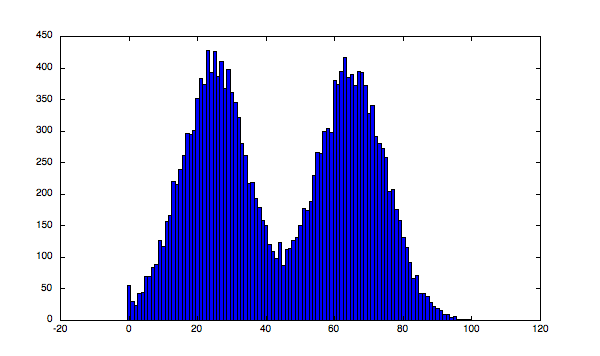
30) If the correlation coefficient (r) between scores in a math test and amount of physical exercise by a student is 0.86, what percentage of variability in math test is explained by the amount of exercise?

A) 86%  
B) 74%  
C) 14%  
D) 26%

Solution: (B)

The % variability is given by r2, the square of the correlation coefficient. This value represents the fraction of the variation in one variable that may be explained by the other variable. Therefore % variability explained would be 0.862.

31) Which of the following is true about below given histogram?

A) Above histogram is unimodal

B) Above histogram is bimodal

C) Given above is not a histogram

D) None of the above

Solution: (B)

The above histogram is bimodal. As we can see there are two values for which we can see peaks in the histograms indicating high frequencies for those values. Therefore the histogram is bimodal.

32) Consider a regression line y=ax+b, where a is the slope and b is the intercept. If we know the value of the slope then by using which option can we always find the value of the intercept?

A) Put the value (0,0) in the regression line True

B) Put any value from the points used to fit the regression line and compute the value of b False

C) Put the mean values of x & y in the equation along with the value a to get b False

D) None of the above can be used False

Solution: (C)

In case of ordinary least squares regression, the line would always pass through the mean values of x and y. If we know one point on the line and the value of slope, we can easily find the intercept.

33) What happens when we introduce more variables to a linear regression model?

A) The r squared value may increase or remain constant, the adjusted r squared may increase or decrease.

B) The r squared may increase or decrease while the adjusted r squared always increases.

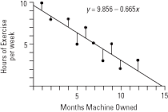
C) Both r square and adjusted r square always increase on the introduction of new variables in the model.

D) Both might increase or decrease depending on the variables introduced.

Solution: (A)

The R square always increases or at least remains constant because in case of ordinary least squares the sum of square error never increases by adding more variables to the model. Hence the R squared does not decrease. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

34) In a scatter diagram, the vertical distance of a point above or below regression line is known as \_\_\_\_ ?

A) Residual  
B) Prediction Error  
C) Prediction  
D) Both A and B  
E) None of the above

Solution: (D)

The lines as we see in the above plot are the vertical distance of points from the regression line. These are known as the residuals or the prediction error.

35) In univariate linear least squares regression, relationship between correlation coefficient and coefficient of determination is \_\_\_\_\_\_ ?

A) Both are unrelated False

B) The coefficient of determination is the coefficient of correlation squared True

C) The coefficient of determination is the square root of the coefficient of correlation False

D) Both are same F

Solution: (B)

The coefficient of determination is the R squared value and it tells us the amount of variability of the dependent variable explained by the independent variable. This is nothing but correlation coefficient squared. In case of multivariate regression the r squared value represents the ratio of the sum of explained variance to the sum of total variance.

36) What is the relationship between significance level and confidence level?

A) Significance level = Confidence level  
B) Significance level = 1- Confidence level  
C) Significance level = 1/Confidence level  
D) Significance level = sqrt (1 – Confidence level)

Solution: (B)

Significance level is 1-confidence interval. If the significance level is 0.05, the corresponding confidence interval is 95% or 0.95. The significance level is the probability of obtaining a result as extreme as, or more extreme than, the result actually obtained when the null hypothesis is true. The confidence interval is the range of likely values for a population parameter, such as the population mean. For example, if you compute a 95% confidence interval for the average price of an ice cream, then you can be 95% confident that the interval contains the true average cost of all ice creams.

The significance level and confidence level are the complementary portions in the normal distribution.

37) [True or False] Suppose you have been given a variable V, along with its mean and median. Based on these values, you can find whether the variable “V” is left skewed or right skewed for the condition

mean(V) > median(V)

A) True  
B) False

Solution: (B)

Since, its no where mentioned about the type distribution of the variable V, we cannot say whether it is left skewed or right skewed for sure.

38) The line described by the linear regression equation (OLS) attempts to \_\_\_\_ ?

A) Pass through as many points as possible.

B)  Pass through as few points as possible

C) Minimize the number of points it touches

D) Minimize the squared distance from the points

Solution: (D)

The regression line attempts to minimize the squared distance between the points and the regression line. By definition the ordinary least squares regression tries to have the minimum sum of squared errors. This means that the sum of squared residuals should be minimized. This may or may not be achieved by passing through the maximum points in the data. The most common case of not passing through all points and reducing the error is when the data has a lot of outliers or is not very strongly linear.

39) We have a linear regression equation ( Y = 5X +40) for the below table.

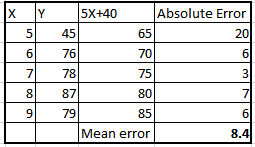
|  |  |
| --- | --- |
| X | Y |
| 5 | 45 |
| 6 | 76 |
| 7 | 78 |
| 8 | 87 |
| 9 | 79 |

Which of the following is a MAE (Mean Absolute Error) for this linear model?

A) 8.4  
B) 10.29  
C) 42.5  
D) None of the above

Solution: (A)

To calculate the mean absolute error for this case, we should first calculate the values of y with the given equation and then calculate the absolute error with respect to the actual values of y. Then the average value of this absolute error would be the mean absolute error. The below table summarises these values.



40) A regression analysis between weight (y) and height (x) resulted in the following least squares line: y = 120 + 5x. This implies that if the height is increased by 1 inch, the weight is expected to

A) increase by 1 pound  
B) increase by 5 pound  
C) increase by 125 pound  
D) None of the above

Solution:  (B)

Looking at the equation given y=120+5x. If the height is increased by 1 unit, the weight will increase by 5 pounds. Since 120 will be the same in both cases and will go off in the difference.

41) [True or False] Pearson captures how linearly dependent two variables are whereas Spearman captures the monotonic behaviour of the relation between the variables.

A)TRUE

B) FALSE

Solution: (A)

The statement is true. Pearson correlation evaluated the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

The spearman evaluates a monotonic relationship. A monotonic relationship is one where the variables change together but not necessarily at a constant rate.

End Notes

I hope you had fun solving the questions and they did make you scratch your head sometime. Please share your thoughts on the above topics and also your feedback.

We shall be happy to incorporate your ideas in further articles and tests. Also, one question might have multiple approaches and the solution above might show just one. I have tried to be descriptive with the solutions but feel free to investigate further in case of doubts using the comments below.

1. Name and explain few methods/techniques used in Statistics for analyzing the data?

Answer:

Arithmetic Mean:  
It is the important technique in statistics Arithmetic Mean can also be called an average. It is the number or the quantity obtained by summing two or more numbers/variables and then dividing the sum by the number of numbers/variables.

Median:  
Median is also a way of finding the average of a group of data points. It’s the middle number of a set of numbers. There are two possibilities, the data points can be an odd number group or it can be en even number group.  
If the group is odd, arrange the numbers in the group from smallest to largest. The median will be the one which is exactly sitting in the middle, with an equal number on either side of it. If the group is even, arrange the numbers in order and pick the two middle numbers and add them then divide by 2. It will be the median number of that set.

Mode:  
The mode is also one of the types for finding the average. A mode is a number, which occurs most frequently in a group of numbers. Some series might not have any mode; some might have two modes which is called bimodal series.

In the study of statistics, the three most common ‘averages’ in statistics are Mean, Median and Mode.

Standard Deviation (Sigma):  
Standard Deviation is a measure of how much your data is spread out in statistics.

Regression:  
Regression is an analysis in[statistical modeling](https://www.educba.com/course/machine-learning-r-module-1-fundamentals-probability-statistics/). It’s a statistical process for measuring the relationships among the variables; it determines the strength of the relationship between one variable and a series of other changing independent variables.

2. Explain about statistics branches?

Answer:  
The two main branches of statistics are descriptive statistics and inferential statistics.

Descriptive statistics: Descriptive statistics summarizes the data from a sample using indexes such as mean or standard deviation.

Descriptive Statistics, methods include displaying, organizing and describing the data.

Inferential Statistics: Inferential Statistics draws the conclusions from data that are subject to random variation such as observation errors and sample variation.

3. List all the other models work with statistics to analyze the data?

Answer:  
Statistics along with [Data Analytics analyzes](https://www.educba.com/data-analytics-vs-business-analytics/) the data and help business to make good decisions. Predictive ‘Analytics’ and ‘Statistics’ are useful to analyze current data and historical data to make predictions about future events.

4. List the fields, where statistic can be used?

Answer:  
Statistics can be used in many research fields. Below are the lists of files in which statistics can be used

1 Science

2 [Technology](https://www.educba.com/technology-to-improve-productivity/)

3 Business

4 Biology

5 Computer Science

6 Chemistry

7 It aids in [decision making](https://www.educba.com/course/decision-making-using-marginal-costing/)

8 Provides comparison

9 Explains action that has taken place

10 Predict the future outcome

11Estimate of unknown quantities.

5. What is a linear regression in statistics?

Answer:  
[Linear regression](https://www.educba.com/course/linear-regression-modeling-using-spss/) is one of the statistical techniques used in a [predictive analysis](https://www.educba.com/predictive-analytics-vs-statistics/), in this technique will identify the strength of the impact that the independent variables show on deepened variables.

6. What is a Sample in Statistics and list the sampling methods?

Answer:  
In a Statistical study, a Sample is nothing but a set of or a portion of collected or processed data from a statistical population by a structured and defined procedure and the elements within the sample are known as a sample point.

Below are the 4 sampling methods:

Cluster Sampling: IN cluster sampling method the population will be divided into groups or clusters.

Simple Random: This sampling method simply follows the pure random division.

Stratified: In stratified sampling, the data will be divided into groups or strata.

Systematical: Systematical sampling method picks every kth member of the population.

7. What is P- value and explain it?

Answer:  
When we execute a hypothesis test in statistics, a p-value helps us in determine the significance of our results. These Hypothesis tests are nothing but to test the validity of a claim that is made about a population. A null hypothesis is a situation when the hypothesis and the specified population is with no significant difference due to sampling or experimental error.

8. What is Data Science and what is the relationship between Data science and Statistics?

Answer:  
Data Science is simply data-driven science, it involves the interdisciplinary field of automated scientific methods, algorithms, systems, and process to extracts the insights and knowledge from data in any form, either structured or unstructured. Data Science and Data mining have similarities, both abstracts the useful information from data.

Data Sciences include Mathematical Statistics along with Computer science and Applications. By combing aspects of statistics, visualization, applied mathematics, computer science Data Science is turning the vast amount of data into insights and knowledge.

Statistics is one of the main components of the Data Science. Statistics is a branch of mathematics commerce with the collection, analysis, interpretation, organization, and presentation of data.

9. What is correlation and covariance in statistics?

Answer:  
Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.

Correlation: Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

Covariance: In covariance two items vary together and it’s a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

Some questions on Linear and Logistic regression frequently asked in the interviews:

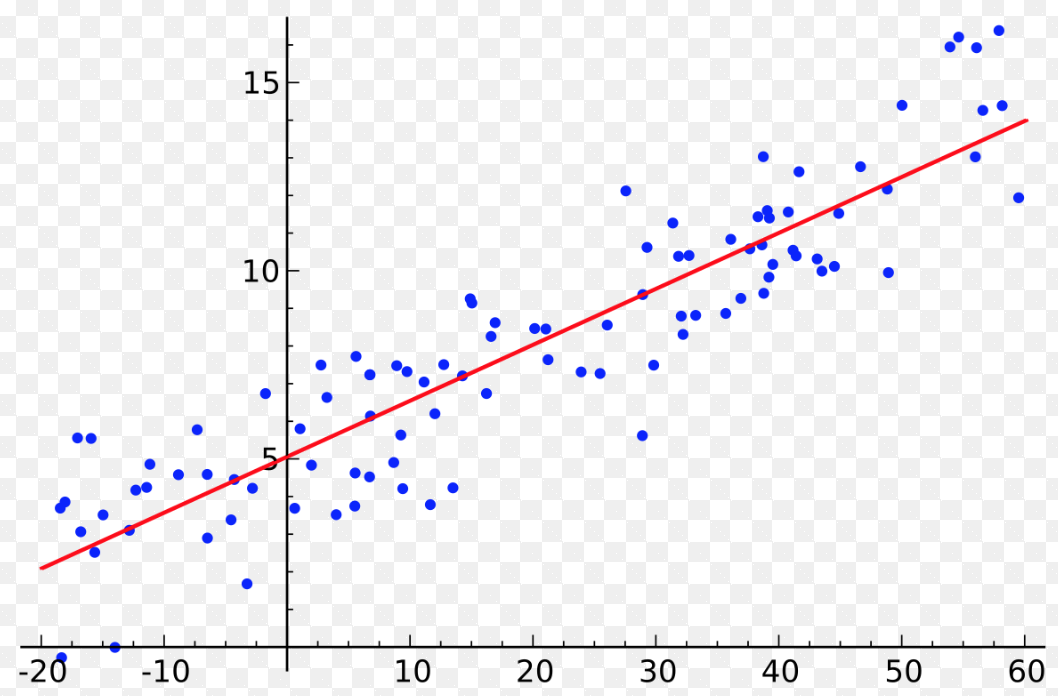
1 Explain Linear Regression in Layman’s terms

2 What is Linear Regression and Logistic regression?

3 What is the role of Linear Regression in EDA (Exploratory Data Analysis)

4 How do you know which regression model you should use?

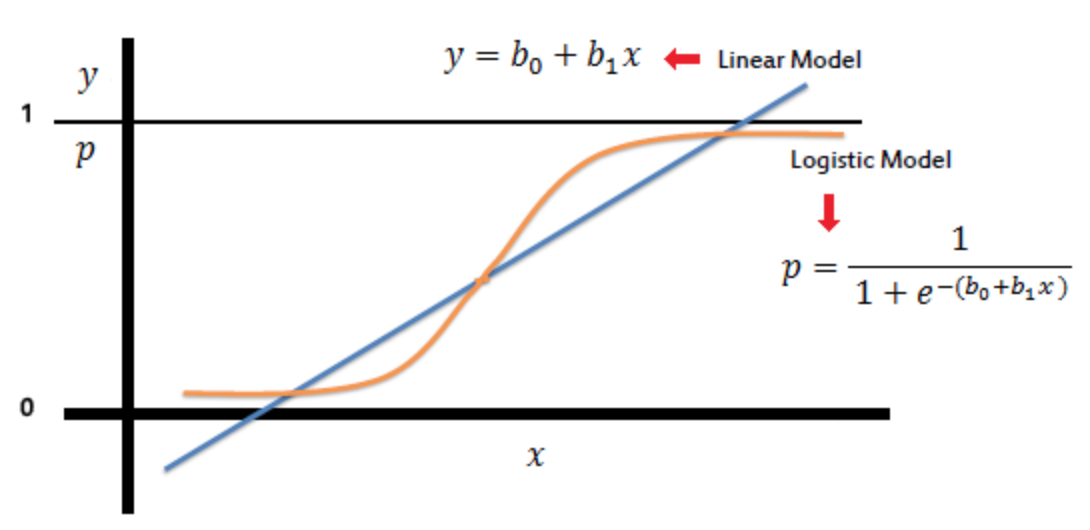
5 (Given a Dataset) Analyze this dataset and give me a model that can predict this response variable.



Source: Wikipedia

The cost of one pen is x$. The cost of ten pens is 10x$ . This is the most classic layman’s form of linear regression. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable. In our pen example, c=0, y is the cost of pens and x is the number of pens. If we know the unit cost of one pen b we can calculate the cost of any number of pens. A complex form of linear regression is used in housing price predictions.

For any scenario based problem in an interview, it is an easy mistake to start with a complex ML Algorithm. Most interviewee’s make the mistake of starting with something that the problem resembles to. They may start with neural networks or SVMs. ALWAYS start with linear/logistic regression if possible. This helps you level set on the most basic benchmark performancefor the solution. Approach that question like a [programming interview](https://medium.com/acing-ai/approach-to-answering-ai-data-science-interview-coding-questions-3a27199a6bea) where you start with a benchmark and you proceed to a more optimized solution.



Source: [Logistic regression](https://www.saedsayad.com/logistic_regression.htm)

Linear regression is used for continuous targets while logistic regression is used for binary targets as sigmoid curve in the logistic model forces the features to either 0 or 1.

This topic brings to light two types of supervised learning algorithms: Classification (Logistic Regression) and Regression (Linear Regression). Exploratory data analysis is something every Data Scientist does on a per project basis. It is the analysis done before the application of a predictive model on a dataset. During EDA we find different characteristics of the data set, plot graphs and decide on some features we will be using. Also we are getting an idea about how to prepare your data, what challenges might be there (feature selection), model measurement, what should we use Accuracy, Precision-Recall, ROC AUC or Mean Squared Error and Pearson Correlation? This usually should start by using Linear/Logistic Regression type of models. The complex models will become clearer if the simple models are benchmarked revealing all characteristics of the dataset. Sometimes, linear/logistic regression might be give you great benchmark results. For the MINST dataset logistic regression generates 95% accuracy which is a great outcome for preliminary analysis. For selecting between different regression models there are some articles which explain the process [here](https://towardsdatascience.com/selecting-the-best-machine-learning-algorithm-for-your-regression-problem-20c330bad4ef).

It is worth investing time in knowing more deeply about Linear and Logistic regression. It is important to know the derivations for both [Linear](http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf) and [Logistic regression](http://www.win-vector.com/blog/2011/09/the-simpler-derivation-of-logistic-regression/). They are used as the base to a lot of ML models and hence, most interviewers want to dig deeper into these basic models. Neural networks for example, each neuron which is a building block of the network is a logistic regression. Each neuron has the input, the weights, the bias we do a dot product to all of that and then apply a non linear function.

Let’s dive in:

1 What’s the trade-off between bias and variance?

2 What is gradient descent?

3 Explain over- and under-fitting and how to combat them?

4 How do you combat the curse of dimensionality?

5 What is regularization, why do we use it, and give some examples of common methods?

6 Explain Principal Component Analysis (PCA)?

7 Why is ReLU better and more often used than Sigmoid in Neural Networks?

8 What is data normalization and why do we need it? I felt this one would be important to highlight. Data normalization is very important preprocessing step, used to rescale values to fit in a specific range to assure better convergence during backpropagation. In general, it boils down to subtracting the mean of each data point and dividing by its standard deviation. If we don’t do this then some of the features (those with high magnitude) will be weighted more in the cost function (if a higher-magnitude feature changes by 1%, then that change is pretty big, but for smaller features it’s quite insignificant). The data normalization makes all features weighted equally.

9 Explain dimensionality reduction, where it’s used, and it’s benefits?Dimensionality reduction is the process of reducing the number of feature variables under consideration by obtaining a set of principal variables which are basically the important features. Importance of a feature depends on how much the feature variable contributes to the information representation of the data and depends on which technique you decide to use. Deciding which technique to use comes down to trial-and-error and preference. It’s common to start with a linear technique and move to non-linear techniques when results suggest inadequate fit. Benefits of dimensionality reduction for a data set may be: (1) Reduce the storage space needed (2) Speed up computation (for example in machine learning algorithms), less dimensions mean less computing, also less dimensions can allow usage of algorithms unfit for a large number of dimensions (3) Remove redundant features, for example no point in storing a terrain’s size in both sq meters and sq miles (maybe data gathering was flawed) (4) Reducing a data’s dimension to 2D or 3D may allow us to plot and visualize it, maybe observe patterns, give us insights (5) Too many features or too complex a model can lead to overfitting.

10 How do you handle missing or corrupted data in a dataset? You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value. In Pandas, there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

11 Explain this clustering algorithm? I wrote a popular article on the [The 5 Clustering Algorithms Data Scientists Need to Know](https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68" \t "_blank) explaining all of them in detail with some great visualizations.

12 How would you go about doing an Exploratory Data Analysis (EDA)?The goal of an EDA is to gather some insights from the data before applying your predictive model i.e gain some information. Basically, you want to do your EDA in a coarse to fine manner. We start by gaining some high-level global insights. Check out some imbalanced classes. Look at mean and variance of each class. Check out the first few rows to see what it’s all about. Run a pandas df.info() to see which features are continuous, categorical, their type (int, float, string). Next, drop unnecessary columns that won’t be useful in analysis and prediction. These can simply be columns that look useless, one’s where many rows have the same value (i.e it doesn’t give us much information), or it’s missing a lot of values. We can also fill in missing values with the most common value in that column, or the median. Now we can start making some basic visualizations. Start with high-level stuff. Do some bar plots for features that are categorical and have a small number of groups. Bar plots of the final classes. Look at the most “general features”. Create some visualizations about these individual features to try and gain some basic insights. Now we can start to get more specific. Create visualizations between features, two or three at a time. How are features related to each other? You can also do a PCA to see which features contain the most information. Group some features together as well to see their relationships. For example, what happens to the classes when A = 0 and B = 0? How about A = 1 and B = 0? Compare different features. For example, if feature A can be either “Female” or “Male” then we can plot feature A against which cabin they stayed in to see if Males and Females stay in different cabins. Beyond bar, scatter, and other basic plots, we can do a PDF/CDF, overlayed plots, etc. Look at some statistics like distribution, p-value, etc. Finally it’s time to build the ML model. Start with easier stuff like Naive Bayes and Linear Regression. If you see that those suck or the data is highly non-linear, go with polynomial regression, decision trees, or SVMs. The features can be selected based on their importance from the EDA. If you have lots of data you can use a Neural Network. Check ROC curve. Precision, Recall

13 How do you know which Machine Learning model you should use?While one should always keep the “no free lunch theorem” in mind, there are some general guidelines. I wrote an article on how to select the proper regression model [here](https://towardsdatascience.com/selecting-the-best-machine-learning-algorithm-for-your-regression-problem-20c330bad4ef). This [cheatsheet](https://www.google.com/search?tbs=simg:CAESqQIJvnrCwg_15JjManQILEKjU2AQaBAgUCAoMCxCwjKcIGmIKYAgDEijqAvQH8wfpB_1AH_1hL1B_1YH6QKOE6soyT-TJ9A0qCipKKoo0TS0NL0-GjA_15sJ-3A24wpvrDVRc8bM3x0nrW3Ctn6tFeYFLpV7ldtVRVDHO-s-8FnDFrpLKzC8gBAwLEI6u_1ggaCgoICAESBOmAAdwMCxCd7cEJGogBChsKCGRvY3VtZW502qWI9gMLCgkvbS8wMTVidjMKGAoGbnVtYmVy2qWI9gMKCggvbS8wNWZ3YgoXCgVtdXNpY9qliPYDCgoIL20vMDRybGYKGwoIcGFyYWxsZWzapYj2AwsKCS9tLzAzMHpmbgoZCgdwYXR0ZXJu2qWI9gMKCggvbS8waHdreQw&q=choose+ml+algorithm&tbm=isch&sa=X&ved=0ahUKEwi-js_8nNbaAhWB5YMKHUTLCEMQsw4INg&biw=1855&bih=990" \l "imgrc=vnrCwg_5JjNUcM:" \t "_blank) is also fantastic!

14 Why do we use convolutions for images rather than just FC layers?This one was pretty interesting since it’s not something companies usually ask. As you would expect, I got this question from a company focused on Computer Vision. This answer has 2 parts to it. Firstly, convolutions preserve, encode, and actually use the spatial information from the image. If we used only FC layers we would have no relative spatial information. Secondly, Convolutional Neural Networks (CNNs) have a partially built-in translation in-variance, since each convolution kernel acts as it’s own filter/feature detector.

15 What makes CNNs translation invariant? As explained above, each convolution kernel acts as it’s own filter/feature detector. So let’s say you’re doing object detection, it doesn’t matter where in the image the object is since we’re going to apply the convolution in a sliding window fashion across the entire image anyways.

16 Why do we have max-pooling in classification CNNs? Again as you would expect this is for a role in Computer Vision. Max-pooling in a CNN allows you to reduce computation since your feature maps are smaller after the pooling. You don’t lose too much semantic information since you’re taking the maximum activation. There’s also a theory that max-pooling contributes a bit to giving CNNs more translation in-variance. Check out this great video from Andrew Ng on the [benefits of max-pooling](https://www.coursera.org/learn/convolutional-neural-networks/lecture/hELHk/pooling-layers).

17 Why do segmentation CNNs typically have an encoder-decoder style / structure? The encoder CNN can basically be thought of as a feature extraction network, while the decoder uses that information to predict the image segments by “decoding” the features and upscaling to the original image size.

18 What is the significance of Residual Networks? The main thing that residual connections did was allow for direct feature access from previous layers. This makes information propagation throughout the network much easier. One very interesting [paper](https://arxiv.org/abs/1605.06431) about this shows how using local skip connections gives the network a type of ensemble multi-path structure, giving features multiple paths to propagate throughout the network.

19 What is batch normalization and why does it work? Training Deep Neural Networks is complicated by the fact that the distribution of each layer’s inputs changes during training, as the parameters of the previous layers change. The idea is then to normalize the inputs of each layer in such a way that they have a mean output activation of zero and standard deviation of one. This is done for each individual mini-batch at each layer i.e compute the mean and variance of that mini-batch alone, then normalize. This is analogous to how the inputs to networks are standardized. How does this help? We know that normalizing the inputs to a network helps it learn. But a network is just a series of layers, where the output of one layer becomes the input to the next. That means we can think of any layer in a neural network as the first layer of a smaller subsequent network. Thought of as a series of neural networks feeding into each other, we normalize the output of one layer before applying the activation function, and then feed it into the following layer (sub-network).

20 How would you handle an imbalanced dataset? I have an [article](https://towardsdatascience.com/7-practical-deep-learning-tips-97a9f514100e) about this! Check out #3 :)

21 Why would you use many small convolutional kernels such as 3x3 rather than a few large ones? This is very well explained in the [VGGNet paper](https://arxiv.org/pdf/1409.1556.pdf" \t "_blank). There are 2 reasons: First, you can use several smaller kernels rather than few large ones to get the same receptive field and capture more spatial context, but with the smaller kernels you are using less parameters and computations. Secondly, because with smaller kernels you will be using more filters, you’ll be able to use more activation functions and thus have a more discriminative mapping function being learned by your CNN.

22 Do you have any other projects that would be related here? Here you’ll really draw connections between your research and their business. Is there anything you did or any skills you learned that could possibly connect back to their business or the role you are applying for? It doesn’t have to be 100% exact, just somehow related such that you can show that you will be able to directly add lots of value.

23 Explain your current masters research? What worked? What didn’t? Future directions? Same as the last question!

1: What is machine learning?

In answering this question, try to show your understand of the broad applications of machine learning, as well as how it fits into AI. Put it into your own words, but convey your understanding that machine learning is a form of AI that automates data analysis to enable computers to learn and adapt through experience to do specific tasks without explicit programming.

2: What is your training in machine learning and what types of hands-on experience do you have?

Your answer to this question will depend on your [training in machine learning](https://www.simplilearn.com/big-data-and-analytics/machine-learning-certification-training-course). Be sure to emphasize any direct projects you’ve completed as part of your education. Don’t fail to mention any additional experience that you have including certifications and how they have prepared you for your role in the machine learning field.

3: What is deep learning?

This might or might not apply to the job you’re going after, but your answer will help to show you know more than just the technical aspects of machine learning. Deep learning is a subset of machine learning. It refers to using multi-layered neural networks to process data in increasingly complex ways, enabling the software to train itself to perform tasks like speech and image recognition through exposure to these vast amounts of data. Thus the machine undergoes continual improvement in the ability to recognize and process information. Layers of neural networks stacked on top of each for use in deep learning are called deep neural networks.

4: How do deductive and inductive machine learning differ?

Deductive machine learning starts with a conclusion, then learns by deducing what is right or wrong about that conclusion. Inductive machine learning starts with examples from which to draw conclusions.

5: How do you choose an algorithm for a classification problem?

The answer depends on the degree of accuracy needed and the size of the training set. If you have a small training set, you can use a low variance/high bias classifier. If your training set is large, you will want to choose a high variance/low bias classifier.

6: How do bias and variance play out in machine learning?

Both bias and variance are errors. Bias is an error due to flawed assumptions in the learning algorithm. Variance is an error resulting from too much complexity in the learning algorithm.

7: What are some methods of reducing dimensionality?

You can reduce dimensionality by combining features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction.

8: How do classification and regression differ?

Classification predicts group or class membership. Regression involves predicting a response. Classification is the better technique when you need a more definite answer.

9: What is supervised versus unsupervised learning?

Supervised learning is a process of machine learning in which outputs are fed back into a computer for the software to learn from for more accurate results the next time. With supervised learning, the “machine” receives initial training to start. In contrast, unsupervised learning means a computer will learn without initial training.

10: What is kernel SVM?

Kernel SVM is the abbreviated version of kernel support vector machine. Kernel methods are a class of algorithms for pattern analysis and the most common one is the kernel SVM.

11. What is decision tree classification?

A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.

12: What is a recommendation system?

Anyone who has used Spotify or shopped at Amazon will recognize a recommendation system: It’s an information filtering system that predicts what a user might want to hear or see based on choice patterns provided by the user.

How to Ace Machine Learning Job Interviews

Machine Learning Interview Questions: Algorithms/Theory

These algorithms questions will test your grasp of the theory behind machine learning.

Q1- What’s the trade-off between bias and variance?

Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you’re using. This can lead to the model underfitting your data, making it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set.

Variance is error due to too much complexity in the learning algorithm you’re using. This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead your model to overfit the data. You’ll be carrying too much noise from your training data for your model to be very useful for your test data.

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if you make the model more complex and add more variables, you’ll lose bias but gain some variance — in order to get the optimally reduced amount of error, you’ll have to tradeoff bias and variance. You don’t want either high bias or high variance in your model.

Q2- What is the difference between supervised and unsupervised machine learning?

Supervised learning requires training labeled data. For example, in order to do classification (a supervised learning task), you’ll need to first label the data you’ll use to train the model to classify data into your labeled groups. Unsupervised learning, in contrast, does not require labeling data explicitly.

Q3- How is KNN different from k-means clustering?

K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this really means is that in order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled point into (thus the nearest neighbor part). K-means clustering requires only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between different points.

The critical difference here is that KNN needs labeled points and is thus supervised learning, while k-means doesn’t — and is thus unsupervised learning.

Q4- Explain how a ROC curve works.

The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It’s often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).

Q5- Define precision and recall.

Recall is also known as the true positive rate: the amount of positives your model claims compared to the actual number of positives there are throughout the data. Precision is also known as the positive predictive value, and it is a measure of the amount of accurate positives your model claims compared to the number of positives it actually claims. It can be easier to think of recall and precision in the context of a case where you’ve predicted that there were 10 apples and 5 oranges in a case of 10 apples. You’d have perfect recall (there are actually 10 apples, and you predicted there would be 10) but 66.7% precision because out of the 15 events you predicted, only 10 (the apples) are correct.

Q6- What is Bayes’ Theorem? How is it useful in a machine learning context?

Bayes’ Theorem gives you the posterior probability of an event given what is known as prior knowledge.

Mathematically, it’s expressed as the true positive rate of a condition sample divided by the sum of the false positive rate of the population and the true positive rate of a condition. Say you had a 60% chance of actually having the flu after a flu test, but out of people who had the flu, the test will be false 50% of the time, and the overall population only has a 5% chance of having the flu. Would you actually have a 60% chance of having the flu after having a positive test?

Bayes’ Theorem says no. It says that you have a (.6 \* 0.05) (True Positive Rate of a Condition Sample) / (.6\*0.05)(True Positive Rate of a Condition Sample) + (.5\*0.95) (False Positive Rate of a Population)  = 0.0594 or 5.94% chance of getting a flu.

Bayes’ Theorem is the basis behind a branch of machine learning that most notably includes the Naive Bayes classifier. That’s something important to consider when you’re faced with machine learning interview questions.

Q7- Why is “Naive” Bayes naive?

Despite its practical applications, especially in text mining, Naive Bayes is considered “Naive” because it makes an assumption that is virtually impossible to see in real-life data: the conditional probability is calculated as the pure product of the individual probabilities of components. This implies the absolute independence of features — a condition probably never met in real life.

As a Quora commenter put it whimsically, a Naive Bayes classifier that figured out that you liked pickles and ice cream would probably naively recommend you a pickle ice cream.

Q8- Explain the difference between L1 and L2 regularization.

L2 regularization tends to spread error among all the terms, while L1 is more binary/sparse, with many variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a Laplacean prior on the terms, while L2 corresponds to a Gaussian prior.

Q9- What’s your favorite algorithm, and can you explain it to me in less than a minute?

This type of question tests your understanding of how to communicate complex and technical nuances with poise and the ability to summarize quickly and efficiently. Make sure you have a choice and make sure you can explain different algorithms so simply and effectively that a five-year-old could grasp the basics!

Q10- What’s the difference between Type I and Type II error?

Don’t think that this is a trick question! Many machine learning interview questions will be an attempt to lob basic questions at you just to make sure you’re on top of your game and you’ve prepared all of your bases.

Type I error is a false positive, while Type II error is a false negative. Briefly stated, Type I error means claiming something has happened when it hasn’t, while Type II error means that you claim nothing is happening when in fact something is.

A clever way to think about this is to think of Type I error as telling a man he is pregnant, while Type II error means you tell a pregnant woman she isn’t carrying a baby.

Q11- What’s a Fourier transform?

A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. Or as this [more intuitive tutorial](https://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/) puts it, given a smoothie, it’s how we find the recipe. The Fourier transform finds the set of cycle speeds, amplitudes and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain — it’s a very common way to extract features from audio signals or other time series such as sensor data.

Q12- What’s the difference between probability and likelihood?

Q13- What is deep learning, and how does it contrast with other machine learning algorithms?

Deep learning is a subset of machine learning that is concerned with neural networks: how to use backpropagation and certain principles from neuroscience to more accurately model large sets of unlabelled or semi-structured data. In that sense, deep learning represents an unsupervised learning algorithm that learns representations of data through the use of neural nets.

Q14- What’s the difference between a generative and discriminative model?

A generative model will learn categories of data while a discriminative model will simply learn the distinction between different categories of data. Discriminative models will generally outperform generative models on classification tasks.

Q15- What cross-validation technique would you use on a time series dataset?

Instead of using standard k-folds cross-validation, you have to pay attention to the fact that a time series is not randomly distributed data — it is inherently ordered by chronological order. If a pattern emerges in later time periods for example, your model may still pick up on it even if that effect doesn’t hold in earlier years!

You’ll want to do something like forward chaining where you’ll be able to model on past data then look at forward-facing data.

fold 1 : training [1], test [2]

fold 2 : training [1 2], test [3]

fold 3 : training [1 2 3], test [4]

fold 4 : training [1 2 3 4], test [5]

fold 5 : training [1 2 3 4 5], test [6]

Q16- How is a decision tree pruned?

Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning.

Reduced error pruning is perhaps the simplest version: replace each node. If it doesn’t decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

Q17- Which is more important to you– model accuracy, or model performance?

This question tests your grasp of the nuances of machine learning model performance! Machine learning interview questions often look towards the details. There are models with higher accuracy that can perform worse in predictive power — how does that make sense?

Well, it has everything to do with how model accuracy is only a subset of model performance, and at that, a sometimes misleading one. For example, if you wanted to detect fraud in a massive dataset with a sample of millions, a more accurate model would most likely predict no fraud at all if only a vast minority of cases were fraud. However, this would be useless for a predictive model — a model designed to find fraud that asserted there was no fraud at all! Questions like this help you demonstrate that you understand model accuracy isn’t the be-all and end-all of model performance.

Q18- What’s the F1 score? How would you use it?

The F1 score is a measure of a model’s performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst. You would use it in classification tests where true negatives don’t matter much.

Q19- How would you handle an imbalanced dataset?

An imbalanced dataset is when you have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if you have no predictive power on the other category of data! Here are a few tactics to get over the hump:

1- Collect more data to even the imbalances in the dataset.

2- Resample the dataset to correct for imbalances.

3- Try a different algorithm altogether on your dataset.

What’s important here is that you have a keen sense for what damage an unbalanced dataset can cause, and how to balance that.

Q20- When should you use classification over regression?

Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points. You would use classification over regression if you wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)

Q21- Name an example where ensemble techniques might be useful.

Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data).

You could list some examples of ensemble methods, from bagging to boosting to a “bucket of models” method and demonstrate how they could increase predictive power.

Q22- How do you ensure you’re not overfitting with a model?

This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.

There are three main methods to avoid overfitting:

1- Keep the model simpler: reduce variance by taking into account fewer variables and parameters, 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.

Q23- What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

You would first split the dataset into training and test sets, or perhaps use cross-validation techniques to further segment the dataset into composite sets of training and test sets within the data. You should then implement a choice selection of performance metrics: here is a fairly [comprehensive list](http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/). You could use measures such as the F1 score, the accuracy, and the confusion matrix. What’s important here is to demonstrate that you understand the nuances of how a model is measured and how to choose the right performance measures for the right situations.

Q24- How would you evaluate a logistic regression model?

A subsection of the question above. You have to demonstrate an understanding of what the typical goals of a logistic regression are (classification, prediction etc.) and bring up a few examples and use cases.

Q25- What’s the “kernel trick” and how is it useful?

The Kernel trick involves kernel functions that can enable in higher-dimension spaces without explicitly calculating the coordinates of points within that dimension: instead, kernel functions compute the inner products between the images of all pairs of data in a feature space. This allows them the very useful attribute of calculating the coordinates of higher dimensions while being computationally cheaper than the explicit calculation of said coordinates. Many algorithms can be expressed in terms of inner products. Using the kernel trick enables us effectively run  algorithms in a high-dimensional space with lower-dimensional data.

Machine Learning Interview Questions: Programming

These machine learning interview questions test your knowledge of programming principles you need to implement machine learning principles in practice. Machine learning interview questions tend to be technical questions that test your logic and programming skills: this section focuses more on the latter.

Q26- How do you handle missing or corrupted data in a dataset?

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.

In Pandas, there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

Q27- Do you have experience with Spark or big data tools for machine learning?

You’ll want to get familiar with the meaning of big data for different companies and the different tools they’ll want. Spark is the big data tool most in demand now, able to handle immense datasets with speed. Be honest if you don’t have experience with the tools demanded, but also take a look at job descriptions and see what tools pop up: you’ll want to invest in familiarizing yourself with them.

Q28- Pick an algorithm. Write the psuedo-code for a parallel implementation.

This kind of question demonstrates your ability to think in parallelism and how you could handle concurrency in programming implementations dealing with big data. Take a look at pseudocode frameworks such as [Peril-L](http://www.eng.utah.edu/~cs4960-01/lecture4.pdf) and visualization tools such as [Web Sequence Diagrams](https://www.websequencediagrams.com/) to help you demonstrate your ability to write code that reflects parallelism.

Q29- What are some differences between a linked list and an array?

An array is an ordered collection of objects. A linked list is a series of objects with pointers that direct how to process them sequentially. An array assumes that every element has the same size, unlike the linked list. A linked list can more easily grow organically: an array has to be pre-defined or re-defined for organic growth. Shuffling a linked list involves changing which points direct where — meanwhile, shuffling an array is more complex and takes more memory.

Q30- Describe a hash table.

A hash table is a data structure that produces an associative array. A key is mapped to certain values through the use of a hash function. They are often used for tasks such as database indexing.

Q31- Which data visualization libraries do you use? What are your thoughts on the best data visualization tools?

What’s important here is to define your views on how to properly visualize data and your personal preferences when it comes to tools. Popular tools include R’s ggplot, Python’s seaborn and matplotlib, and tools such as Plot.ly and Tableau.

Machine Learning Interview Questions: Company/Industry Specific

These machine learning interview questions deal with how to implement your general machine learning knowledge to a specific company’s requirements. You’ll be asked to create case studies and extend your knowledge of the company and industry you’re applying for with your machine learning skills.

Q32- How would you implement a recommendation system for our company’s users?

A lot of machine learning interview questions of this type will involve implementation of machine learning models to a company’s problems. You’ll have to research the company and its industry in-depth, especially the revenue drivers the company has, and the types of users the company takes on in the context of the industry it’s in.

Q33- How can we use your machine learning skills to generate revenue?

This is a tricky question. The ideal answer would demonstrate knowledge of what drives the business and how your skills could relate. For example, if you were interviewing for music-streaming startup Spotify, you could remark that your skills at developing a better recommendation model would increase user retention, which would then increase revenue in the long run.

The startup metrics Slideshare linked above will help you understand exactly what performance indicators are important for startups and tech companies as they think about revenue and growth.

Q34- What do you think of our current data process?

Machine Learning Interview Questions: General Machine Learning Interest

This series of machine learning interview questions attempts to gauge your passion and interest in machine learning. The right answers will serve as a testament for your commitment to being a lifelong learner in machine learning.

Q35- What are the last machine learning papers you’ve read?

Keeping up with the latest scientific literature on machine learning is a must if you want to demonstrate interest in a machine learning position. This overview of [deep learning in Nature](http://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf) by the scions of deep learning themselves (from Hinton to Bengio to LeCun) can be a good reference paper and an overview of what’s happening in deep learning — and the kind of paper you might want to cite.

Q36- Do you have research experience in machine learning?

Related to the last point, most organizations hiring for machine learning positions will look for your formal experience in the field. Research papers, co-authored or supervised by leaders in the field, can make the difference between you being hired and not. Make sure you have a summary of your research experience and papers ready — and an explanation for your background and lack of formal research experience if you don’t.

Q37- What are your favorite use cases of machine learning models?

The Quora thread above contains some examples, such as decision trees that categorize people into different tiers of intelligence based on IQ scores. Make sure that you have a few examples in mind and describe what resonated with you. It’s important that you demonstrate an interest in how machine learning is implemented.

Q38- How would you approach the “Netflix Prize” competition?

The Netflix Prize was a famed competition where Netflix offered $1,000,000 for a better collaborative filtering algorithm. The team that won called BellKor had a 10% improvement and used an ensemble of different methods to win. Some familiarity with the case and its solution will help demonstrate you’ve paid attention to machine learning for a while.

Q39- Where do you usually source datasets?

Machine learning interview questions like these try to get at the heart of your machine learning interest. Somebody who is truly passionate about machine learning will have gone off and done side projects on their own, and have a good idea of what great datasets are out there. If you’re missing any, check out [Quandl](https://www.quandl.com/" \t "_blank) for economic and financial data, and [Kaggle’s Datasets](https://www.kaggle.com/datasets) collection for another great list.

Q40- How do you think Google is training data for self-driving cars?

Machine learning interview questions like this one really test your knowledge of different machine learning methods, and your inventiveness if you don’t know the answer. Google is currently using [recaptcha](https://www.google.com/recaptcha" \t "_blank) to source labelled data on storefronts and traffic signs. They are also building on training data collected by Sebastian Thrun at GoogleX — some of which was obtained by his grad students driving buggies on desert dunes!

Q41- How would you simulate the approach AlphaGo took to beat Lee Sidol at Go?

AlphaGo beating Lee Sidol, the best human player at Go, in a best-of-five series was a truly seminal event in the history of machine learning and deep learning. The Nature paper above describes how this was accomplished with “Monte-Carlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play.”

1. What do you understand by Machine Learning?

Answer:  
Machine learning is an application of [artificial intelligence](https://www.educba.com/artificial-intelligence-vs-business-intelligence/) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

2.Give an example that explains Machine Leaning in industry.

Answer:  
Robots are replacing humans in many areas. It is because robots are programmed such that they can perform the task based on data they gather from sensors. They learn from the data and behaves intelligently.

Let us move to the next Machine Learning Interview Questions.

3. What are the different Algorithm techniques in Machine Learning?

Answer:  
The different types of Algorithm techniques in Machine Learning are as follows:  
• 1 [Reinforcement Learning](https://www.educba.com/supervised-learning-vs-reinforcement-learning/)  
• 2 Supervised Learning  
• 3 Unsupervised Learning  
• 4 Semi-supervised Learning  
• 5 Transduction  
• 6 Learning to Learn

4. What is the difference between supervised and unsupervised machine learning?

Answer:  
This is the basic Machine Learning Interview Questions asked in an interview. A Supervised learning is a process where it requires training labeled data While Unsupervised learning it doesn’t require data labeling.

5. What is the function of Unsupervised Learning?

Answer:  
The function of Unsupervised Learning are as below:  
• 1 Find clusters of the data of the data  
• 2 Find low-dimensional representations of the data  
• 3 Find interesting directions in data  
• 4 Interesting coordinates and correlations  
• 5 Find novel observations

6. What is the function of Supervised Learning?

Answer:  
The function of Supervised Learning are as below:  
• 1 Classifications  
• 2 Speech recognition  
• 3 [Regression](https://www.educba.com/course/regression-modeling-minitab-module-1-basic/)  
• 4 Predict time series  
• 5 Annotate strings

7. What are the advantages of Naive Bayes?

Answer:  
The advantages of Naive Bayes are:  
• 1 The classifier will converge quicker than discriminative models  
• 2 It cannot learn the interactions between features

Let us move to the next Machine Learning Interview Questions.

8. What are the disadvantages of Naive Bayes?

Answer:  
The disadvantages of Naive Bayes are:  
• 1 It is because the problem arises for continuous features  
• 2 It makes a very strong assumption on the shape of your data distribution  
• 3 It can also happen because of data scarcity

9. Why is naive Bayes so naive?

Answer:  
Naive Bayes is so naive because it assumes that all of the features in a dataset are equally important and independent.

10. What is Overfitting in Machine Learning?

Answer:  
This is the popular Machine Learning Interview Questions asked in an interview. Overfitting in Machine Learning is defined as when a statistical model describes random error or noise instead of underlying relationship or when a model is excessively complex.

11. What are the conditions when Overfitting happens?

Answer:  
One of the important reason and possibility of overfitting is because the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model.

12. How can you avoid overfitting?

Answer:  
We can avoid overfitting by using:  
• 1 Lots of data  
• 2 Cross-validation

Part 2 – Machine Learning Interview Questions (Advanced)

Let us now have a look at the advanced Machine Learning Interview Questions.

13. What are the five popular algorithms for Machine Learning?

Answer:  
Below is the list of five popular algorithms of Machine Learning:  
• 1 Decision Trees  
• 2 Probabilistic networks  
• 3 Nearest Neighbor  
• 4 Support vector machines  
• 5 Neural Networks

14. What are the different use cases where machine learning algorithms can be used?

Answer:  
The different use cases where machine learning algorithms can be used are as follows:  
• 1 [Fraud Detection](https://www.educba.com/fraud-detection-analytics/)  
• 2 [Face detection](https://www.educba.com/course/face-detection-in-python/)  
• 3 Natural language processing  
• 4 [Market Segmentation](https://www.educba.com/how-to-perform-market-segmentation/)  
• 5 Text Categorization  
• 6 Bioinformatics

Let us move to the next Machine Learning Interview Questions.

15. What are parametric models and Non-Parametric models?

Answer:  
Parametric models are those with a finite number of parameters and to predict new data, you only need to know the parameters of the model.  
Non Parametric models are those with an unbounded number of parameters, allowing for more flexibility and to predict new data, you need to know the parameters of the model and the state of the data that has been observed.

16 .What are the three stages to build the hypotheses or model in machine learning?

Answer:  
This is the frequently asked Machine Learning Interview Questions in an interview. The three stages to build the hypotheses or model in machine learning are:  
1. Model building  
2. Model testing  
3. Applying the model

17. What is Inductive Logic Programming in Machine Learning (ILP)?

Answer:  
Inductive Logic Programming (ILP) is a subfield of machine learning which uses logical [programming](https://www.educba.com/course/online-c-programming-course-training/) representing background knowledge and examples.

18. What is the difference between classification and regression?

Answer:  
The difference between classification and regression are as follows:  
• 1 Classification is about identifying group membership while regression technique involves predicting a response.  
• 2 Classification and Regression techniques are related to prediction  
• 3 Classification predicts the belonging to a class whereas regression predicts the value from a continuous set  
• 4 Classification technique is preferred over regression when the results of the model need to return the belongingness of data points in a dataset with specific explicit categories

Let us move to the next Machine Learning Interview Questions.

19. What are the difference between inductive machine learning and deductive machine learning?

Answer:  
The difference between inductive machine learning and deductive machine learning are as follows:  
[machine learning](https://www.educba.com/data-mining-vs-machine-learning/) where the model learns by examples from a set of observed instances to draw a generalized conclusion whereas in deductive learning the model first draws the conclusion and then the conclusion is drawn.

20. What are the advantages decision trees?

Answer:  
The advantages decision trees are:  
• 1 Decision trees are easy to interpret  
• 2 Nonparametric  
• 3 There are relatively few parameters to tune

21. What are the disadvantages of decision trees?

Answer:  
Decision trees are prone to be overfit. However, this can be addressed by ensemble methods like random forests or boosted trees.

22. What are the advantages of neural networks?

Answer:  
This is the advanced Machine Learning Interview Questions asked in an interview. Neural networks have led to performance breakthroughs for unstructured datasets such as images, audio, and video. Their incredible flexibility allows them to learn patterns that no other Machine Learning algorithm can learn.

23. What are the disadvantages of neural networks?

Answer:  
Neural Network requires a large amount of training data to converge. It’s also difficult to pick the right architecture, and the internal “hidden” layers are incomprehensible.

24. What is the difference between L1 and L2 regularization?

Answer:  
The difference between L1 and L2 regularization are as follows:  
• 1 L1/Laplace tends to tolerate both large values as well as very small values of coefficients more than L2/Gaussian  
• 2 L1 can yield sparse models while L2 doesn’t  
• 3 L1 and L2 regularization prevents overfitting by shrinking on the coefficients  
• 4 L2 (Ridge) shrinks all the coefficient by the same proportions but eliminates none, while L1 (Lasso) can shrink some coefficients to zero, performing variable selection  
• 5 L1 is the first-moment norm |x1-x2| that is simply the absolute dıstance between two points where L2 is second-moment norm corresponding to Euclidean Distance that is |x1-x2|^2.  
• 6 L2 regularization tends to spread error among all the terms, while L1 is more binary/sparse