Ace the upcoming Data Science Interview

You can't anticipate every question an interviewer will ask. However, there are many **critical questions** that you can prepare before the interview.

Our hiring partners have helped us curate a set of interview questions on key skills, which will help you prepare better for the data science job roles.

Filters

1. What is XGBoost?

Basic ML

Hint?

This was majorly asked since the candidate had used an ET model in capstone

2. How do you deploy a model to cloud

Intermediate ML

The workflow can be broken down into following basic steps:

- Training a machine learning model on a local system

- Wrapping the inference logic into a flask application

- Using Docker to containerize the flask application

- Hosting the docker container on an AWS ec2 instance and consuming the web-service

3. How will you make models out of the tweets for the pharma company

Advanced ML

Hint?

NLP: Text analytics model

4. Make 4 segments (product category, competitors etc) and identify which medicine a doctor is likely to recommend

Intermediate ML

Hint?

Python : Unsupervised learning

5. Working of ensemble methods such as bagging, boosting, random forest.

Intermediate ML

Hint?

Start with explaining what ensemble techniques are and focus on the "why" and "when" one should use ensemble techniques. https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f https://towardsdatascience.com/gradient-boosting-machines-gbms-the-eli5-way-c4a21b2e2b0a

6. What is clustering and KNN?

Basic ML

k-Means Clustering is an unsupervised learning algorithm that is used for clustering whereas KNN is a supervised learning algorithm used for classification.

The “k” in k-means denotes the number of clusters you want to have in the end. If k = 5, you will have 5 clusters on the data set. “k” in K-Nearest Neighbors is the number of neighbours it checks. It is supervised because you are trying to classify a point based on the known classification of other points.

7. What is bagging and boosting?

Basic ML

Bagging and Boosting decrease the variance of your single estimate as they combine several estimates from different models. So the result may be a model with higher stability.

Bagging is used when the goal is to reduce the variance of a decision tree classifier. Here the objective is to create several subsets of data from the training sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees. As a result, we get an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree classifier.

Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree. When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly. This process converts weak learners into a better performing model

8. What is ADA boosting?

Basic ML

Ada-boost is an ensemble classifier. It combines a weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with a selection of training set at every iteration and assigning the right amount of weight in the final voting, we can have good accuracy score for the overall classifier.

9. Explain Gradient boosting and Extreme Gradient Boosting?

Basic ML

XGBoost stands for Extreme Gradient Boosting; it is a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model. It employs a number of nifty tricks that make it exceptionally successful, particularly with structured data.

The most important are:

1) computing second-order gradients, i.e. second partial derivatives of the loss function (similar to Newton's method), which provides more information about the direction of gradients and how to get to the minimum of our loss function. While regular gradient boosting uses the loss function of our base model (e.g. decision tree) as a proxy for minimizing the error of the overall model, XGBoost uses the 2nd order derivative as an approximation.

2) And advanced regularization (L1 & L2), which improves model generalization.

XGBoost has additional advantages: training is very fast and can be parallelized/distributed across clusters.

10. What is Bootstrap sampling?

Basic ML

Bootstrap Sampling is a method that involves drawing of sample data repeatedly with replacement from a data source to estimate a population parameter.

11. What to be done on the dataset if the assumptions are not met?

Intermediate ML

1. If you create a scatter plot of values for x and y and see that there is not a linear relationship between the two variables, then one can do the following:

- Apply a nonlinear transformation to the independent and/or dependent variable. e.g. log, square root, or reciprocal of the independent and/or dependent variable

- Add another independent variable to the model.

2. If residuals are not independent then one can do the following:

- For positive serial correlation, consider adding lags of the dependent and/or independent variable to the model.

- For negative serial correlation, check to make sure that none of your variables is overdifferenced.

- For seasonal correlation, consider adding seasonal dummy variables to the model

3. If Residuals do not have constant variance, then one can do the following:

- Transform the dependent variable

- Use weighted regression

4. If Residuals are not normally distributed, then one can do the following:

- First, verify that any outliers aren't having a huge impact on the distribution. If there are outliers present, make sure that they are real values and that they aren't data entry errors

- Next, you can apply a nonlinear transformation to the independent and/or dependent variable. e.g. log, square root, or the reciprocal of the independent and/or dependent variable

12. How to apply ML Algorithms in Mfg/Production Environment ?

Intermediate ML

1. Specify Performance Requirements (This may be as accurate or false positives or whatever metrics are important to the business)

2. Separate Prediction Algorithm From Model Coefficients

2a. Select or Implement The Prediction Algorithm

2b. Serialize Your Model Coefficients

3. Develop Automated Tests For Your Model

4. Develop Back-Testing and Now-Testing Infrastructure

5. Challenge Then Trial Model Updates (For example, perhaps you set up a grid or random search of model hyperparameters that runs every night and spits out new candidate models)

13. Difference between Classification and Linear Regression?

Basic ML

1. Fundamentally, classification is about predicting a label and regression is about predicting a quantity.

i.e. Classification is the task of predicting a discrete class label while Regression is the task of predicting a continuous quantity

2. Classification predictions can be evaluated using accuracy, whereas regression predictions cannot.

Regression predictions can be evaluated using root mean squared error, whereas classification predictions cannot

predictions cannot.

3. A regression algorithm can predict a discrete value which is in the form of an integer quantity

A classification algorithm can predict a continuous value if it is in the form of a class label probability

14. Which model to use to check whether a patient is diabetic or not?

Basic ML

Classification algorithm such as Logistic regression, Random forest etc

15. Explain missing values and outlier treatment

Basic ML

16. What is logistic regression? The output for logistic regression?

Basic ML

a. Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems.

b. Log likelihood – This is the log likelihood of the final model

c. Number of obs – This is the number of observations that were used in the analysis

d. LR chi2(3) – This is the likelihood ratio (LR) chi-square test. The number in the parenthesis indicates the number of degrees of freedom

e. Prob > chi2 – This is the probability of obtaining the chi-square statistic given that the null hypothesis is true. In this case, the model is statistically significant because the p-value is less than .000.

f. Pseudo R2 – This is the pseudo R-squared.

17. What is Ensemble techniques and it's working? some models?

Basic ML

A group of weak learners coming together to form a strong learner, thus increasing the accuracy of any Machine Learning model is called an ensemble model

Simple Ensemble Techniques: Hard Voting Classifier, Averaging, Weighted Averaging

Advanced Ensemble Techniques: Stacking, Bagging (Randomforest) and Pasting Boosting(Adaboost, XGB etc)

18. What is Decision tree and Random forest?

Basic ML

- A decision tree is a supervised machine learning algorithm that can be used for both classification and regression problems. A decision tree is simply a series of sequential decisions made to reach a specific result

- Random Forest is a tree-based machine learning algorithm that leverages the power of multiple (randomly created) decision trees for making decisions. i.e. The Random Forest Algorithm combines the output of multiple (randomly created) Decision Trees to generate the final output.

- Random Forest is suitable for situations when we have a large dataset, and interpretability is not a major concern. Decision trees are much easier to interpret and understand. Since a random forest combines multiple decision trees, it becomes more difficult to interpret.

- The decision tree model gives high importance to a particular set of features. But the random forest chooses features randomly during the training process.

19. How to deal with underfitting and overfitting

Basic ML

Handling Overfitting:

Cross-validation

This is done by splitting your dataset into ‘test’ data and ‘train’ data. Build the model using the ‘train’ set. The ‘test’ set is used for in-time validation.

Regularization

This is a form of regression, that regularizes or shrinks the coefficient estimates towards zero. This technique discourages learning a more complex model

Early stopping

When training a learner with an iterative method, you stop the training process before the final iteration. This prevents the model from memorizing the dataset.

Pruning

This technique applies to decision trees.

Pre-pruning: Stop 'growing' the tree earlier before it perfectly classifies the training set.

Post-pruning: Allows the tree to 'grow', perfectly classify the training set and then post prune the tree.

Dropout

This is a technique where randomly selected neurons are ignored during training.

Regularize the weights

Handling Underfitting:

Get more training data

Increase the size or number of parameters in the model

Increase the complexity of the model

Increasing the training time, until cost function is minimised

20. What is bias variance tradeoff

Basic ML

The goal of any supervised machine learning algorithm is to achieve low bias(the difference between the average prediction of our model and the correct value which we are trying to predict) and low variance(variability of model prediction for a given data point or a value which tells us spread of our data).

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand, if our model has a large number of parameters then it's going to have high variance and low bias.

Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

So we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance.

21. How will you explain machine learning to a 5 year old.

Intermediate ML

Just like a human, a computer can learn from three sources.

One is Observing what others did in similar situations. The other is observing a situation and trying to come up with the best possible logic on the spot to decide/conclude. The third is learning from previous mistakes/success. These three methods correspond to three branches of Machine learning, Supervised, Unsupervised and Reinforcement learning respectively.

- In Supervised Learning, a computer can tell what word in a sentence is the name of a city, given it is shown example sentences which may or may not contain names of cities and every occurrence of a city name is tagged in these examples.

- Unsupervised is where we ask the computer to make decisions based on raw data attributes and a set of measurable quantities. Some examples would include asking a computer to come up with localities in a dataset where Lat-Long of the house is given. It would use Lat Long to find distances and form localities of house.

- The third type of learning is Reinforcement Learning. This is a method in which computer starts with making random decisions, and then learns based on errors it makes and successes it encounters as it goes. A recent discovery was an algorithm which could play many different arcade games after learning the correct/wrong moves. These algorithms would start by making a lot of failures in the beginning and then get better as they go.

22. What do you do in data exploration?

Basic ML

23. You are given a data set on cancer detection. You've build a classification model and achieved an accuracy of 96%. Why shouldn't you be happy with your model performance? What can you do about it?

Intermediate ML

24. You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?

Advanced ML

25. You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?

Intermediate ML

26. How is kNN different from kmeans clustering?

Basic ML

27. After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he's true? Without losing any information, can you still build a better model?

Intermediate ML

28. When is Ridge regression favorable over Lasso regression?

Basic ML

29. While working on a data set, how do you select important variables? Explain your methods.

Basic ML

30. What is the difference between covariance and correlation?

Intermediate ML

31. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?

Basic ML

32. You've got a data set to work having p (no. of variable) > n (no. of observation). Why is (Ordinary Least Squares) OLS is bad option to work with? Which techniques would be best to use? Why?

Advanced ML

33. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn't. How ?

Intermediate ML

34. You are given a data set consisting of variables having more than 30% missing values? Let's say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?

Basic ML

35. People who bought this, also bought...' recommendations seen on amazon is a result of which algorithm?

Intermediate ML

36. What do you understand by Type I vs Type II error ?

Basic ML

37. You have been asked to evaluate a regression model based on R^2, adjusted R^2 and tolerance. What will be your criteria?

Basic ML

38. Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?

Basic ML

39. When does regularization becomes necessary in Machine Learning?

Basic ML

40. What do you understand by Bias Variance trade off?

Basic ML

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

Basic ML

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

Basic ML

- Repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model

- Example: repeatedly draw different samples from training data, fit a linear regression to each new sample, and then examine the extent to which the resulting fit differ

- most common are: cross-validation and the bootstrap

cross-validation: random sampling with no replacement, bootstrap: random sampling with replacement

- cross-validation: evaluating model performance, model selection (select the appropriate level of flexibility)

- bootstrap: mostly used to quantify the uncertainty associated with a given estimator or statistical learning method

43. Is it better to have too many false positives, or too many false negatives? Explain.

Basic ML

False-positive and false-negative are two problems we have to deal with while evaluating a mode.

In medical, a false positive can lead to unnecessary treatment and a false negative can lead to a false diagnostic, which is very serious since the disease has been ignored.

However, we can minimize the errors by collecting more information, considering other variables, adjusting the sensitivity (true positive rate) and specificity (true negative rate) of the test, or conducting the test multiple times.

Even so, it is still hard since reducing one type of error means increasing the other type of error. Sometimes, one type of error is more preferable than the other one, so data scientists will have to evaluate the consequences of the errors and make a decision

44. What is selection bias, why is it important and how can you avoid it

Basic ML

Selection bias occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution.

How to avoid selection biases

Mechanisms for avoiding selection biases include:

- Using random methods when selecting subgroups from populations.

- Ensuring that the subgroups selected are equivalent to the population at large in terms of their key characteristics (this method is less of a protection than the first since typically the key characteristics are not known).

45. Differentiate between univariate, bivariate and multivariate analysis.

Basic ML

Univariate statistics summarize only one variable at a time.

Bivariate statistics compare two variables.

Multivariate statistics compare more than two variables.

46. What is the difference between Cluster and Systematic Sampling?

Basic ML

Systematic sampling and cluster sampling are both statistical measures used by researchers, analysts, and marketers to study samples of a population.

Systematic sampling involves selecting fixed intervals from the larger population to create the sample.

Cluster sampling divides the population into groups, then takes a random sample from each cluster.

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

Intermediate ML

Let us take an example of a medical field where:

A false positive = person is considered as sick but actually is healthy

A false negative = person is considered as healthy but is actually sick

What does it mean?

False-positive cases lead to overspending due to unnecessary care and damaging the health of an otherwise healthy person due to unnecessary side effects of the therapy.

A false negative case means that your patients get sicker or die.

In this case, both false positive and false negatives are equally important since it concerns a person's life

48. Explain Lasso regression

Basic ML

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters)

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminate from the model. Larger penalties result in coefficient values closer to zero, which is ideal for producing simpler models.

49. Explain Gradient Descent Algorithm

Intermediate ML

Gradient descent is an optimization algorithm that's used when training a machine learning model.

It's based on a convex function and tweaks its parameters iteratively to minimize a given cost function to its local minimum.

You start by defining the initial parameter's values and from there gradient descent uses calculus to iteratively adjust the values so they minimize the given cost-function (where a gradient measures how much the output of a function changes if you change the inputs a little bit.)

50. How machine learning is deployed in real world scenarios?

Advanced ML

AWS or Azure instances with python jobs that run with either manual schedules, or automated to trigger on receiving say new data. These are usually a suite of services that constitute a deployment environment of such models.

Storage - model needs to be stored somewhere (pickle or joblib or specific model object). Either s3 on aws or blob in azure.

Computing instance - Computing environment that contains python and is enabled to communicate to every platform that is relevant to the deployment context.

Job scheduler - Devops is the norm now. Automated pipelines that procure data, process, load/retrain/predict with the packaged model.

Final layer - either BI tools like tableau, qilkview etc or sql/nosql databases or excel reports

- ### 51. What is cosine similarity?

  Intermediate ML

  Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, the higher the cosine similarity.

52. How to implement Tensorflow?

Intermediate ML

The usual workflow of running a program in TensorFlow is as follows:

Build a computational graph, this can be any mathematical operation TensorFlow supports.

Initialize variables, to compile the variables defined previously

Create a session, this is where the magic starts!

Run graph in session, the compiled graph is passed to the session, which starts its execution.

Close session, shut down the session.