

Q. How to improve model performance?

A. Follow these techniques:

1. Use Validation methods
2. Add more data
3. Apply feature engineering techniques(Normalization, Imputation etc)
4. Compare Multiple algorithms
5. Hyperparameter Tuning

Q. Standardization vs log transformation?

A. Standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables. Typically, to standardize variables, you calculate the mean and standard deviation for a variable. Log transformation is a data transformation method in which it replaces each variable x with a $\log(x)$. Log-transform decreases skew in some distributions, especially with large outliers. But, it may not be useful as well if the original distributed is not skewed. Also, log transform may not be applied to some cases (negative values), but standardization is always applicable (except $\sigma=0$).

Q. Object detection?

A. Object detection is a computer vision technique for locating instances of objects in images or videos. Object detection algorithms typically leverage machine learning or deep learning to produce meaningful results.

Q. Data cleaning steps?

A. Step 1: Remove duplicate or irrelevant observations. Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations.

Step 2: Fix structural errors.

Step 3: Filter unwanted outliers.

Step 4: Handle missing data.

Step 5: Validate your data if it's appropriate according to problem statement

Q. Upsampling and downsampling methods?

A. In a classification task, there is a high chance for the algorithm to be biased if the dataset is imbalanced. An imbalanced dataset is one in which the number of samples in one class is very higher or lesser than the number of samples in the other class.

To counter such imbalanced datasets, we use a technique called up-sampling and down-sampling.

In up-sampling, we randomly duplicate the observations from the minority class in order to reinforce its signal. The most common way is to resample with replacement.

In down-sampling, we randomly remove the observations from the majority class. Thus after up-sampling or down-sampling, the dataset becomes balanced with same number of observations in each class.

Q. hypothesis testing?

A. Hypothesis testing is defined as the process of choosing hypotheses for a particular probability distribution, based on observed data. We use this to test whether a hypothesis can be accepted or not.

Q. Overfitting and Underfitting? It's relation with bias and variance?

A. Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points. Underfitting refers to a model that can neither model the training data nor generalize to new data.

When the bias is more the prediction of the model is very far from the actual value. It means that the model is not having capacity to generalize the distribution of the data. This is underfitting. You will have to increase the complexity of the model so that it can better generalize the data distribution.

On the other hand when the variance of the model is more, then the values predicted by the model are highly spread from the expected value predicted by the model(not the actual value). This is overfitting. The model is highly complicated and needs to be made simple. Otherwise noise and outliers can take a great toll on the model.

Q. Why scaling useful?

A. So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

Q. A fair six-sided die is rolled 6 times. What is the probability of getting all outcomes as unique?

A. Prob = $\frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{2}{6} \times \frac{1}{6} = \frac{6!}{6^6} = 0.015$

Q. how does random Random Forest works? How is it different from decision trees?

A. A decision tree combines some decisions, whereas a random forest combines several decision trees. Thus, it is a long process, yet slow. Whereas, a decision tree is fast and operates easily on large data sets, especially the linear one. The random forest model needs rigorous training.

Q. How can you convert timestamps to date time in MySQL?

A. We can convert the timestamp to date time with the help of FROM_UNIXTIME () function.

Q. In experimental design, is it necessary to do randomization? If yes, why?

A. Yes, in experimental design, randomization is required. Randomization eliminates biases and ensures that the outcomes are balanced. The sample that is randomly picked is designed to be representative of the population, and it is fairly selected because it is not influenced by the researcher. You can achieve the best cause-effect linkages between the variables by randomizing the experiments.

Q. What is the benefit of weight initialisation in neural networks?

A. The fundamental goal of weight initialization is to prevent exploding or vanishing gradients in layer activation outputs during forward propagation. If either of these issues arises, loss gradients will be either too great or too little, and the network will take longer to converge, if it can at all. If we appropriately set the weights, we will achieve our goal of optimising the loss function in the shortest time possible; otherwise, converging to a minimum via gradient descent will be impossible.

Q. How will you evaluate the performance of a logistic regression model?

A. The confusion matrix can be used to evaluate a logistic regression model. The accuracy, sensitivity, and specificity of the model can be useful indicators of what you want to do with it - focusing on true positives or false negatives. We can also utilize precision and recall to evaluate your model, as well as the f1 score.

CLOUDYML

NOTE:

If you want to learn Data Science from scratch and become interview ready, check out our popular Data science course

<https://ai.cloudymml.com/Learn-Data-Science-from-Scratch>