Company Name - Bridgei2i Analytics
Role - Data Scientist

Q. How would you remove bias?

A. Ways to minimize Bias in ML:

1. Choose the correct learning model. There are two types of learning models, and each has its own pros and cons.
2. Use the right training dataset.
3. Perform data processing mindfully.
4. Monitor real-world performance across the ML lifecycle.
5. Make sure that there are no infrastructural issues.

Q. In what situation would you consider mean over median?

A. If your data contains outliers, then you would typically rather use the median because otherwise the value of the mean would be dominated by the outliers rather than the typical values. In conclusion, if you are considering the mean, check your data for outliers, if any then better choose median.

Q. What is LASSO & RIDGE?

A. Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression. Ridge Regression: In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients. lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

Q. Describe Optimization?

A. Optimization is the problem of finding a set of inputs to an objective function that results in a maximum or minimum function evaluation. It is the challenging problem that underlies many machine learning algorithms, from fitting logistic regression models to training artificial neural networks.

Q. What is standard error of the mean?

A. The standard error of mean tells you how accurate the mean of any given sample from that population is likely to be compared to the true population mean. When the standard error increases, i.e. the means are more spread out, it becomes more likely that any given mean is an inaccurate representation of the true population mean.

Q. ROC Curve?

A. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

Q. Why does overfitting occur?

A. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model

Q. What is ensemble learning?

A. Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one.

Q. What is F1 score?

A. The F1 score is defined as the harmonic mean of precision and recall. As a short reminder, the harmonic mean is an alternative metric for the more common arithmetic mean. It is often useful when computing an average rate. In the F1 score, we compute the average of precision and recall.

Q. What is pickling and unpickling?

A. “Pickling” is the process whereby a Python object hierarchy is converted into a byte stream, and “unpickling” is the inverse operation, whereby a byte stream (from a binary file or bytes-like object) is converted back into an object hierarchy.

Q. What is lambda function?

A. Python Lambda Functions are anonymous function means that the function is without a name. As we already know that the def keyword is used to define a normal function in Python. Similarly, the lambda keyword is used to define an anonymous function in Python.

Q. What is the trade of between bias and variance?

A. Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.

Q. What is Central tendency?

A. Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution.”[2] It aims to provide an accurate description of the entire data. It is the single value that is most typical/representative of the collected data.

Q. Map Reduce?

A. MapReduce facilitates concurrent processing by splitting petabytes of data into smaller chunks, and processing them in parallel on Hadoop commodity servers. In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application.

Q. What is a Pivot Table?

A. A pivot table is a table of grouped values that aggregates the individual items of a more extensive table within one or more discrete categories.
