AMAZON DATA SCIENCE INTERVIEW

What is variance in a model?

Variance in a model refers to how much the model's predictions change when trained on different subsets of the data. It captures the sensitivity of the model to variations in the training data.

High variance means the model is very sensitive to the specific data it was trained on. This results in large fluctuations in predictions when exposed to different datasets, even if they are similar. High variance is typically associated with overfitting.

Low variance means the model's predictions are stable, even when trained on different datasets.

Is a decision tree model best for predicting if a borrower will pay back a personal loan? How would you evaluate performance of the model?

This would depend on the data and model performance.
For instance, decision trees can be a good starting point since it's quite interpretable, handles non-linear relationships and requires minimal pre-processing.
However, in financial datasets the data tends to be imbalanced. So assessing performance on wide variety of classification metrics like precision, recall, f1-score would be important to assess model performance.

From here, the interviewer might have follow up - so make sure you understand metrics very well, specifically - Precision, Recall, F1, AUC-ROC, AUC-PR etc.

@karunt

What would you do if 20% of the 100,000 sold listings are missing square footage data. You want to predict price.

This technique would depend on what cause of missing values are - is it at random, does it mean the listing is 'pending', or 'not ready for sale'. Once we understand what the reason, we can solve it in different ways like -
(a) drop the feature if its not predictive, or if other feature are good proxies.
(a) imputation-mean/median or KNN-imputer etc.
(b) use models that can handle missing values - XGBoost or Random Forest

Get feedback from interviews and be ready to dive into each apporach!

@karunt

What is the difference between XGboost and random forest?

Some differences are -

XGBoost: It's an implementation of **gradient boosting**. XGBoost builds trees **sequentially**, where each new tree tries to correct the errors made by the previous trees. It has Lower bias due to the sequential learning process, but **potentially higher variance** if overfitting occurs. **Regularization** techniques are applied to mitigate overfitting.

Random Forest: It is an example of bagging (Bootstrap Aggregating)where multiple trees are build **independently and in parallel**. **Higher bias** because trees are grown independently, but variance is reduced because of the averaging across many trees. It does not have explicit regularization, but **naturally prevents overfitting** by averaging predictions from multiple trees and using random feature selection.

@karunt

WAS THIS HELPFUL?

Be sure to save it so you
can come back to it later!

@karunt