Decision Tree Interview Questions

(Practice Project)

PW SKILLS

Easy Questions:

1. **What is a leaf node in a Decision Tree?**

Answer: A leaf node is a terminal node in a Decision Tree that contains the final decision or prediction. It does not split further and represents the outcome of the decision path.

2. **What is the main criterion for selecting a split in a Decision Tree?**

Answer: The main criterion for selecting a split in a Decision Tree is to maximize the reduction in impurity, often measured by Gini Impurity, Entropy, or Information Gain.

3. **What is the depth of a Decision Tree?**

Answer: The depth of a Decision Tree is the length of the longest path from the root node to a leaf node. It indicates the number of splits or levels in the tree.

4. **Why is it important to set a maximum depth for a Decision Tree?**

Answer: Setting a maximum depth for a Decision Tree helps prevent overfitting by limiting the tree's complexity. It ensures that the tree does not become too detailed and captures only the relevant patterns in the data.

5. **What is the difference between a binary and a multiway split in a Decision Tree?**

Answer: A binary split divides the data into two groups based on a condition, while a multiway split divides the data into more than two groups. Binary splits are simpler and more commonly used.

6. **How does a Decision Tree handle missing values?**

Answer: Some Decision Tree algorithms can handle missing values by either ignoring them during the split calculation or assigning them to the most frequent category or the median value for numerical features.

7. **What is the role of the root node in a Decision Tree?**

Answer: The root node is the topmost node of a Decision Tree and represents the feature that provides the best split of the data. It is the starting point for making predictions.

Medium Questions:

1. **How can you evaluate the performance of a Decision Tree model?**

Answer: The performance of a Decision Tree model can be evaluated using metrics like accuracy, precision, recall, F1-score for classification tasks, and Mean Squared Error (MSE) or R-squared for regression tasks. Cross-validation can also be used to assess model generalization.

2. **What are surrogate splits in Decision Trees, and when are they used?**

Answer: Surrogate splits are alternative splitting rules used when the primary splitting feature is missing. They provide a backup mechanism, allowing the tree to make a decision even with missing data.

3. Explain the concept of feature importance in Decision Trees.

Answer: Feature importance measures the contribution of each feature in making predictions in a Decision Tree. It is calculated based on the reduction in impurity achieved by each feature across all splits in the tree. Features with higher importance have a greater influence on the model's predictions.

4. How does the CART algorithm handle categorical variables?

Answer: The CART algorithm handles categorical variables by considering all possible splits of the categories and selecting the one that results in the greatest reduction in impurity. For binary splits, it divides the categories into two groups.

5. What is the impact of unbalanced data on a Decision Tree model?

Answer: Unbalanced data can lead to biased Decision Tree models that favor the majority class. This can result in poor performance, especially for the minority class. Techniques like class weighting, oversampling, or undersampling can be used to address this issue.

6. What is a random forest, and how does it relate to Decision Trees?

Answer: A Random Forest is an ensemble learning method that combines multiple Decision Trees to improve model accuracy and reduce overfitting. Each tree in the forest is built on a random subset of the data and features, and the final prediction is made by averaging the predictions of all trees (for regression) or taking a majority vote (for classification).

7. What is the purpose of setting a minimum sample split in a Decision Tree?

Answer: Setting a minimum samples split specifies the minimum number of samples required to split a node. This helps control the growth of the tree, preventing it from splitting on small, potentially noisy subsets of data, thereby reducing the risk of overfitting.

8. What are the common hyperparameters in a Decision Tree algorithm?

Answer: Common hyperparameters in a Decision Tree algorithm include:

max_depth: The maximum depth of the tree. Limiting the depth helps prevent overfitting by controlling the complexity of the tree.

min_samples_split: The minimum number of samples required to split an internal node. A higher value can prevent the tree from splitting on small, potentially noisy data.

min_samples_leaf: The minimum number of samples required to be at a leaf node. This helps control the growth of the tree and can also reduce overfitting.

criterion: The function used to measure the quality of a split. For classification, common criteria are gini (Gini impurity) and entropy (information gain). For regression, mse (mean squared error) or mae (mean absolute error) are used.

max_features: The number of features to consider when looking for the best split. This can be used to reduce overfitting and improve the generalization of the model.

splitter: The strategy used to choose the split at each node. best chooses the best split, while randomly chooses a random split.

min_weight_fraction_leaf: The minimum weighted fraction of the sum total of weights required to be at a leaf node, which can be useful when working with weighted datasets.

max_leaf_nodes: The maximum number of leaf nodes in the tree. Limiting the number of leaf nodes helps in controlling the model complexity.

ccp_alpha: The complexity parameter used for Minimal Cost-Complexity Pruning, which helps in pruning the tree to avoid overfitting by penalizing the number of nodes.

Hard Questions:

1. **How does a Decision Tree handle multi-class classification problems?**

Answer: In multi-class classification problems, a Decision Tree handles multiple classes by recursively splitting the data based on features that reduce impurity the most for all classes. The tree continues to split until each leaf node corresponds to a single class or a mixture of classes.

2. **What is the relationship between Decision Trees and entropy in information theory?**

Answer: In information theory, entropy measures the disorder or uncertainty in a dataset. Decision Trees use entropy to quantify the impurity of a node. The algorithm selects splits that reduce entropy, thereby increasing the information gain, which leads to purer child nodes.

3. **How do you handle imbalanced data sets when training a Decision Tree model?**

Answer: To handle imbalanced datasets, you can use techniques such as adjusting class weights (giving more importance to the minority class), resampling the data (oversampling the minority class or undersampling the majority class), or using specialized algorithms like SMOTE (Synthetic Minority Over-sampling Technique).

4. **What are the limitations of using Decision Trees for regression tasks?**

Answer: Decision Trees for regression tasks can create piecewise constant predictions, which may not capture smooth relationships in the data. They are also sensitive to outliers and can be prone to overfitting, especially with small datasets or deep trees.

5. **How can Decision Trees be used in feature selection?**

Answer: Decision Trees can be used for feature selection by analyzing feature importance scores. Features that contribute most to reducing impurity are considered important, while those with low importance scores can be removed to simplify the model without significantly impacting performance