Decision Tree Classifier And Regressor

Interview Questions:

1. Decision Tree
2. Entropy, Information Gain, Gini Impurity
3. Decision Tree Working For Categorical and Numerical Features
4. What are the scenarios where Decision Tree works well
5. Decision Tree Low Bias And High Variance- Overfitting
6. Hyperparameter Techniques
7. Library used for constructing decision tree
8. Impact of Outliers Of Decision Tree
9. Impact of mising values on Decision Tree
10. Does Decision Tree require Feature Scaling

1. What Are the Basic Assumption?

There are no such assumptions

2. Advantages

1. **Clear Visualization**: The algorithm is simple to understand, interpret and visualize as the idea is mostly used in our daily lives. Output of a Decision Tree can be easily interpreted by humans.
2. **Simple and easy to understand**: Decision Tree looks like simple if-else statements which are very easy to understand.
3. Decision Tree can be used for both classification and regression problems.
4. Decision Tree can handle both continuous and categorical variables.
5. **No feature scaling required**: No feature scaling (standardization and normalization) required in case of Decision Tree as it uses rule based approach instead of distance calculation.
6. **Handles non-linear parameters efficiently:** Non linear parameters don't affect the performance of a Decision Tree unlike curve based algorithms. So, if there is high non-linearity between the independent variables, Decision Trees may outperform as compared to other curve based algorithms.
7. Decision Tree can automatically handle missing values.
8. Decision Tree is usually robust to outliers and can handle them automatically.
9. **Less Training Period**: Training period is less as compared to Random Forest because it generates only one tree unlike forest of trees in the Random Forest.

3. Disadvantages

1. **Overfitting**: This is the main problem of the Decision Tree. It generally leads to overfitting of the data which ultimately leads to wrong predictions. In order to fit the data (even noisy data), it keeps generating new nodes and ultimately the tree becomes too complex to interpret. In this way, it loses its generalization capabilities. It performs very well on the trained data but starts making a lot of mistakes on the unseen data.
2. **High variance:** As mentioned in point 1, Decision Tree generally leads to the overfitting of data. Due to the overfitting, there are very high chances of high variance in the output which leads to many errors in the final estimation and shows high inaccuracy in the results. In order to achieve zero bias (overfitting), it leads to high variance.
3. **Unstable**: Adding a new data point can lead to re-generation of the overall tree and all nodes need to be recalculated and recreated.

4. **Not suitable for large datasets**: If data size is large, then one single tree may grow complex and lead to overfitting. So in this case, we should use Random Forest instead of a single Decision Tree.

4. Whether Feature Scaling is required?

No

6. Impact of outliers?

It is not sensitive to outliers.Since, extreme values or outliers, never cause much reduction in RSS, they are never involved in split. Hence, tree based methods are insensitive to outliers.

Types of Problems it can solve(Supervised)

1. Classification
2. Regression

Practical Implementation

1. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
2. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Performance Metrics

Classification

1. Confusion Matrix
2. Precision,Recall, F1 score

Regression

1. R2,Adjusted R2
2. MSE,RMSE,MAE