Important questions to ace your machine learning interview with an approach to answer

1. Machine Learning Project

Lifecycle:

- - Define the problem
- - Gather and preprocess data
- - Choose a model and train it
- - Evaluate model performance
- - Tune and optimize the model
- - Deploy and maintain the model

2. Supervised vs Unsupervised Learning:

- - Supervised Learning: Uses labeled data for training (e.g., predicting house prices from features).
- - Unsupervised Learning: Uses unlabeled data to find patterns or groupings (e.g., clustering customer segments).

3. Evaluation Metrics for Regression:

- - Mean Absolute Error (MAE)
- - Mean Squared Error (MSE)
- - Root Mean Squared Error (RMSE)
- - R-squared (coefficient of determination)

4. Overfitting and Prevention:

- - Overfitting: Model learns the noise instead of the underlying pattern.
- - Prevention: Use simpler models, cross-validation, regularization.

5. Bias-Variance Tradeoff:

- - Balancing error due to bias (underfitting) and variance (overfitting) to find an optimal model complexity.

6. Cross-Validation:

- Technique to assess model performance by splitting data into multiple subsets for training and validation.

7. Feature Selection Techniques:

- - Filter methods (e.g., correlation analysis)
- - Wrapper methods (e.g., recursive feature elimination)
- - Embedded methods (e.g., Lasso regularization)

8. Assumptions of Linear Regression:

- - Linearity
- - Independence of errors
- - Homoscedasticity (constant variance)
- - No multicollinearity

9. Regularization in Linear Models:

- - Adds a penalty term to the loss function to prevent overfitting by shrinking coefficients.

10. Classification vs Regression:

- - Classification: Predicts a categorical outcome (e.g., class labels).
- - Regression: Predicts a continuous numerical outcome (e.g., house price).

11. Dimensionality Reduction

Algorithms:

- - Principal Component Analysis (PCA)
- - t-Distributed Stochastic Neighbor Embedding (t-SNE)

12. Decision Tree:

- - Tree-like model where internal nodes represent features, branches represent decisions, and leaf nodes represent outcomes.

13. Ensemble Methods:

- Combine predictions from multiple models to improve accuracy (e.g., Random Forest, Gradient Boosting).

14. Handling Missing or Corrupted Data:

- - Imputation (e.g., mean substitution)
- - Removing rows or columns with missing data
- - Using algorithms robust to missing values

15. Kernels in Support Vector Machines (SVM):

- - Linear kernel
- - Polynomial kernel
- - Radial Basis Function (RBF) kernel