Interview Questions & Answers:Machine Learning (Telco Customer Churn Dataset)

1. What is the difference between overfitting and underfitting?
Answer:
Overfitting occurs when a machine learning model learns not only the underlying patterns in the training data but also its noise, leading to poor performance on new, unseen data. The model becomes too complex, capturing unnecessary details.
Underfitting happens when a model is too simple to capture the underlying structure of the data, resulting in poor performance on both training and test data.
Summary Table:

Overfitting	Underfitting
High training accuracy	Low training accuracy
Low test accuracy	Low test accuracy
Model is too complex	Model is too simple

2. Why is feature scaling important?
Answer:
Feature scaling standardizes the range of independent variables or features of data. It is important because many machine learning algorithms (like k-NN, SVM, and neural networks) use distances between data points to make predictions. If features are on different scales, those with larger scales can dominate the distance calculation, leading to biased results. Scaling ensures all features contribute equally to the model's learning process and improves algorithm convergence and performance.

3. What evaluation metric would you choose for imbalanced data and why?
Answer:
For imbalanced datasets (like churn prediction, where most customers do not churn), accuracy can be misleading. Instead, use metrics such as:

- **Precision**: Measures how many of the predicted positive cases are actually positive.
- **Recall**: Measures how many of the actual positive cases are correctly identified.
- **F1-score**: Harmonic mean of precision and recall, providing a balance between the two.
- **AUC-ROC**: Evaluates the model's ability to distinguish between classes across all thresholds.

Why?
These metrics provide a better understanding of model performance on the minority class, which is often more critical in business scenarios like churn prediction.

4. **How does a Random Forest work?**

Answer: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode/mean of predictions from individual trees.
Key steps:

- Bootstrapping: Select random subsets of the training data (with replacement).
- Feature Randomness: For each tree, select a random subset of features at each split.
- Voting/Averaging: Combine predictions from all trees for classification (majority vote) or regression (average).

Advantages:

- Reduces overfitting compared to a single decision tree.
- Handles high-dimensional data well.
- Provides feature importance.

5. What steps would you take if your model had low accuracy?
Answer:
If your model has low accuracy, consider the following steps:

- **Check Data Quality**: Look for missing values, outliers, and inconsistencies.
- **Feature Engineering**: Create new features or select more relevant ones.
- Try Different Models: Experiment with different algorithms (e.g., Random Forest, Gradient Boosting, Neural Networks).
- **Hyperparameter Tuning**: Optimize model parameters using grid search or random search.
- **Address Imbalance**: Use resampling techniques (oversampling minority class, undersampling majority class) or class weights.
- **Cross-Validation**: Use k-fold cross-validation to ensure model robustness.
- **Evaluate with Better Metrics**: Use precision, recall, F1-score, or AUC-ROC for imbalanced data.