**Q1. Difference between Linear Regression and Logistic Regression**

* **Linear Regression**:
  + Used for predicting continuous values.
  + Outputs a numeric value.
  + Assumes a linear relationship between independent and dependent variables.
  + Uses Mean Squared Error (MSE) as the cost function.
* **Logistic Regression**:
  + Used for classification tasks.
  + Outputs probabilities mapped to a range of 0 to 1 using the sigmoid function.
  + Models the relationship between independent variables and a categorical dependent variable.
  + Uses Log Loss (cross-entropy) as the cost function.

**Example**:

* Logistic regression is more appropriate when predicting whether a person will default on a loan (yes/no) rather than predicting the amount of money they might default.

**Q2. Cost Function in Logistic Regression and Optimization**

* **Cost Function**: Logistic regression uses the **log-loss** or cross-entropy cost function, which penalizes predictions that are far from the actual class label.

J(θ)=−1m∑i=1m[yilog⁡(hθ(xi))+(1−yi)log⁡(1−hθ(xi))]J(\theta) = -\frac{1}{m} \sum\_{i=1}^{m} \left[ y\_i \log(h\_\theta(x\_i)) + (1 - y\_i) \log(1 - h\_\theta(x\_i)) \right]

Where:

* + yiy\_i is the actual label (0 or 1).
  + hθ(xi)h\_\theta(x\_i) is the predicted probability.
* **Optimization**: The cost function is minimized using optimization algorithms like **Gradient Descent**.

**Q3. Regularization in Logistic Regression**

* **Concept**:
  + Regularization adds a penalty term to the cost function to prevent overfitting by discouraging overly complex models.
  + Two common types:
    - **L1 Regularization (Lasso)**: Encourages sparsity, driving some coefficients to zero.
    - **L2 Regularization (Ridge)**: Penalizes large coefficients without setting them to zero.
* **How It Helps**:
  + Prevents overfitting by simplifying the model.
  + Ensures that the model generalizes well to unseen data.

**Q4. ROC Curve in Logistic Regression**

* **Definition**:
  + The ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels.
* **Usage**:
  + Evaluates the classification performance of a logistic regression model.
  + The **Area Under the Curve (AUC)** quantifies the model's ability to discriminate between classes:
    - AUC = 1: Perfect model.
    - AUC = 0.5: Random guessing.

**Q5. Techniques for Feature Selection in Logistic Regression**

1. **Univariate Statistical Tests**:
   * E.g., Chi-square tests to identify significant features.
2. **Recursive Feature Elimination (RFE)**:
   * Iteratively removes features to find the subset that improves model performance.
3. **L1 Regularization (Lasso)**:
   * Shrinks less important feature coefficients to zero.
4. **Principal Component Analysis (PCA)**:
   * Reduces dimensionality while retaining most of the variance.

**Benefits**:

* Reduces model complexity.
* Improves interpretability.
* Enhances computational efficiency.

**Q6. Handling Imbalanced Datasets in Logistic Regression**

1. **Resampling Techniques**:
   * **Oversampling the minority class**: E.g., SMOTE (Synthetic Minority Over-sampling Technique).
   * **Undersampling the majority class**.
2. **Class Weights**:
   * Assign higher weights to the minority class in the cost function.
3. **Threshold Tuning**:
   * Adjust the classification threshold to balance sensitivity and specificity.
4. **Evaluation Metrics**:
   * Use metrics like Precision, Recall, F1-Score, and ROC-AUC instead of accuracy.

**Q7. Common Issues and Challenges in Logistic Regression**

1. **Multicollinearity**:
   * Occurs when independent variables are highly correlated.
   * Solution:
     + Use **Variance Inflation Factor (VIF)** to detect multicollinearity.
     + Remove one of the correlated variables or apply dimensionality reduction (e.g., PCA).
2. **Outliers**:
   * Outliers can disproportionately influence the model.
   * Solution:
     + Use robust techniques to identify and handle outliers (e.g., IQR or Z-scores).
3. **Non-linearity**:
   * Logistic regression assumes a linear relationship between features and the log-odds.
   * Solution:
     + Transform features (e.g., polynomial features or interaction terms).
4. **Imbalanced Classes**:
   * Skewed class distribution can lead to biased models.
   * Solution:
     + Use strategies mentioned above for handling imbalanced datasets.
5. **Overfitting**:
   * Too many features or complex models can overfit.
   * Solution:
     + Apply regularization (L1 or L2).