# 1 APRIL

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

Linear Regression: Linear regression is used for predicting a continuous outcome variable (also called dependent or target variable) based on one or more predictor variables (independent variables). It models the relationship between the variables as a linear equation and aims to find the bestfitting line that minimizes the sum of squared errors.

Logistic Regression: Logistic regression is used for predicting the probability of a binary outcome (usually 0 or 1) based on one or more predictor variables. It models the logodds of the probability as a linear combination of predictors and applies the logistic function (sigmoid) to this equation to ensure the output is between 0 and 1. It's commonly used for classification problems.

Example Scenario for Logistic Regression: Suppose you want to predict whether a customer will buy a product (yes/no) based on factors like age, income, and past purchase history. Logistic regression would be more appropriate in this case because the outcome is binary (buy or not buy), making it a classification problem.

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

The cost function used in logistic regression is called the "logistic loss" or "crossentropy loss." It measures the dissimilarity between the predicted probabilities and the actual class labels.

Optimization is typically done using iterative methods like gradient descent. The goal is to find the model parameters (coefficients) that minimize the logistic loss function.

**Q3. Regularization in Logistic Regression:**

Regularization is a technique used to prevent overfitting in logistic regression. It adds a penalty term to the cost function, discouraging large parameter values.

Two common types of regularization in logistic regression are L1 regularization (Lasso) and L2 regularization (Ridge). They control the magnitude of coefficients and can set some coefficients to zero, effectively selecting important features.

**Q4. ROC Curve for Logistic Regression Evaluation:**

ROC (Receiver Operating Characteristic) curve is a graphical representation used to evaluate the performance of a logistic regression model.

It plots the true positive rate (sensitivity) against the false positive rate (1specificity) at various threshold settings.

A good model has an ROC curve that hugs the upperleft corner of the plot, indicating high sensitivity and low false positive rate. The AUC (Area Under the Curve) is also used to summarize the model's performance.

**Q5. Feature Selection in Logistic Regression:**

Common techniques for feature selection include forward selection, backward elimination, and L1 regularization (Lasso).

These techniques help improve model performance by selecting the most relevant features and removing irrelevant or redundant ones, reducing the risk of overfitting and improving model interpretability.

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

Strategies for dealing with class imbalance include oversampling the minority class, undersampling the majority class, and using costsensitive learning.

Oversampling generates more instances of the minority class, while undersampling reduces instances of the majority class. Costsensitive learning assigns different misclassification costs to classes.

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

Multicollinearity among independent variables can be a challenge. It can be addressed by removing one of the correlated variables, using regularization, or applying dimensionality reduction techniques like Principal Component Analysis (PCA).

Other challenges include outliers, data preprocessing, and ensuring the assumptions of logistic regression (linearity, independence of errors, etc.) are met. Address these by robust statistics, data cleaning, and careful model selection.