* Logistic regression used for classification then why its called as regression?

Logistic regression is one of the basic and popular algorithms to solve a classification problem. It is named ‘logistic regression’ because its technique is quite the same as linear regression. The term ‘Logistic’ is taken from the logit function that is used in this method of classification.

* Explain correlation?

Correlation is used for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

Examples like, income and expenditure, demand and supply, etc.

* Can logistic regression used for more than 2 classed?

No, by default logistic regression is a binary classifier, so it cannot be applied to more than 2 classes. However. It can be extended for solving multi-class classification problems (multinomial logistic regression)

* What is true Negative?

When a model correctly predicts the negative class, it is said to be true negative. For example, umpire gives batsman OUT when he is OUT.

* Evaluation metrics for classification?

Metrics like accuracy, precision, recall are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced then other methods like ROC/AUC (receiver operating characteristic curve)(area Under the curve) perform better in evaluating the model performance.

* Evaluation metrics for regression?

There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are:

Mean Squared Error(MSE),

Root Mean Squared Error(RMSE),

Mean Absolute Error(MAE).

* What do you understand by outliers?

An outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset you’re working with.

Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph.

* What is trimming?

“Trimming” data excludes the outlier values from your analysis.

* What is winsorization?

Winsorization is a method of averaging that initially replaces the smallest and largest values with the observations closed to them. This is done to limit the effect of outliers or abnormal extreme values, or outliers, on the calculation.

* Difference between trimming and winsorization?

Trimming and winsorizing are two ways to mitigate the effect of extreme values on your analysis. “trimming” data excludes the outlier values from your analysis. “winsorizing” retains the responses in your basis, but caps numeric outliers so they fall at edge of the main distribution.

* What is True Positive?

When a model correctly predicts the positive class, it is said to be a True positive.

For example, Umpire gives a batsman NOT OUT when he is NOT OUT.

* Describe dimension reduction in machine learning?

Dimension reduction is the process which is used to reduce the number of random variables under considerations.

Dimension reduction can be divided into feature selection and extraction.

* What is false positive?

When a model incorrectly predicts the positive class, It is said to be a false positive. It is also known as “Type I” error.

For example, Umpire gives a Batsman NOT OUT when he is OUT.

* What is False Negative?

When a model incorrectly predicts the negative class. It is said to be a false negative. It is also known as ‘Type II’ error.

For example, Umpire gives a Batsman OUT when he is NOT OUT.

* How do you calculate accuracy in classification using confusion matrix?

Dividing the number of correct predictions by the total number of predictions.

It gives you the overall accuracy of the model.

Accuracy=(TP+TN)/(TP+TN+FP+FN)

* What L1 and L2 regression also called as?

L1 regularization technique is called lasso regression and model which uses L2 is called Ridge regression. The key difference between these two is the penalty term

* What is R2 and adjusted R2? Which is more reliable?

Also called as the coefficient of determination  
R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model r2 don’t prenalise.