1. What is Data Science? List the differences between supervised and unsupervised learning.

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

The data used for analysis can come from many different sources and presented in various formats.

Supervised machine learning: The training data set we feed to algorithm is labeled i.e it includes desired solutions.

Example:In admission prediction problem, we have features such as GRE score, TOEFL score, CGPA, university rating, research and we have to predict chance of admission. These features act as independent variables(X) and our chance of admit is dependent variable(Y). Here, dependent variable is a binary variable which has value 1 for yes or success and 0 for no or failure.

Unsupervised learning: The training data set we feed to algorithm is unlabeled i.e. system tries to learn without a teacher.

Example: Suppose we have images of different types of fruits. We will just provide the input dataset to the model and allow the model to find the patterns from the data. With the help of a suitable algorithm, the model will train itself and divide the fruits into different groups according to the most similar features between them.

2. What is logistic regression?

Logistic Regression is a Supervised Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mail is spam or not etc.

3. How will you deal with the multiclass classification problem using logistic regression?

Multiclass logistic regression is also called multinomial logistic regression. In contrast to the binomial logistic regression, multiclass logistic regression is used to classify the output labels to more than 2 classes.

4. Why is logistic regression very popular/widely used?

Logistic regression is famous because it can convert the values of logits (log-odds), which can range from −∞ to +∞ to a range between 0 and 1. As logistic functions output the probability of occurrence of an event, it can be applied to many real-life scenarios. It is for this reason that the logistic regression model is very popular. Another reason why logistic fairs in comparison to linear regression is that it is able to handle the categorical variables.

5. Why can't linear regression be used instead of logistic regression for classification?

Mathematically linear regression can be explained by,

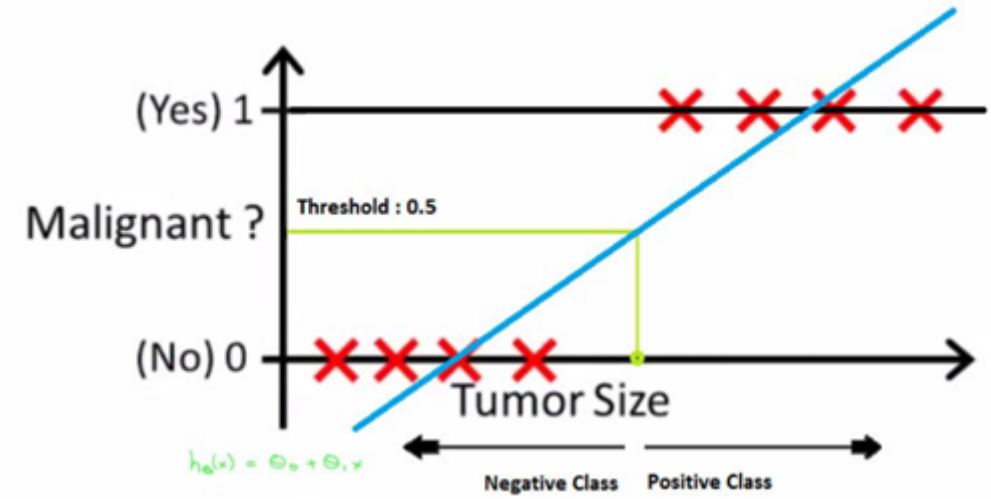
y = mx + c

So, our model will predict the value of y for any value of x. In linear regression, dependent variable is continuous variable. But in logistic regression, Dependent variable is categorical.

Linear regression only works with continuous data. If we want to include linear regression in our classification methods, we'll have to adjust our algorithm a little more. First, we must choose a threshold so that if our projected value is less than the threshold, it belongs to class 1; otherwise, it belongs to class 2.

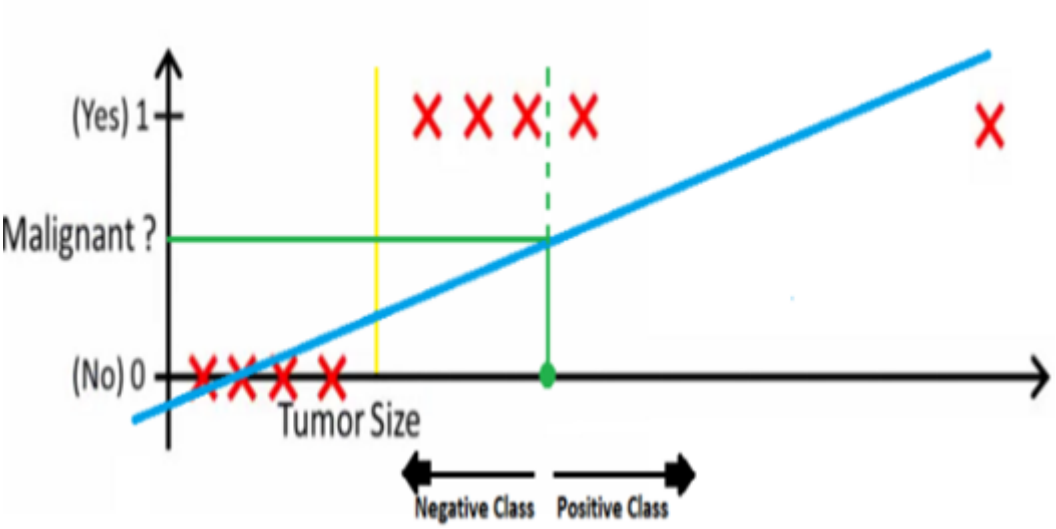
A logistic regression, on the other hand, yields a logistic curve with values confined to 0 and 1. The curve in logistic regression is generated using the natural logarithm of the target variable's "odds," rather than the probability, as in linear regression.

If we use linear regression for classification problem for data with no outliers, it will give the best fit line and using threshold value we may be classify correctly.



Lets assume we have information about tumor size and malignancy. Because this is a classification issue, we can see that all the values will fall between 0 and 1. And, by fitting the best-found regression line and assuming a threshold of 0.5, we can do a very good job with the line.

We can choose a point on the x-axis from which all values on the left side are regarded as negative, and all values on the right side are considered positive.



Even if we fit the best-found regression line, we won't be able to determine any point where we can distinguish classes. It will insert some instances from the positive class into the negative class. The green dotted line (Decision Boundary) separates malignant and benign tumors, however, it should have been a yellow line that clearly separates the positive and negative cases.

If data have any outliers, linear regression won't work for such data and this is where logistics regression comes into the picture.

6. What is the formula for the logistic regression function?

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Where, $\beta 0$ is the y-intercept
$\beta 1$ is the slope of the line
x is the value of the independent variable

p(x)/(1-p(x)) is termed odds, and the left-hand side is called the logit or log-odds function. The odds are the ratio of the chances of success to the chances of failure. As a result, in Logistic Regression, a linear combination of inputs is translated to log(odds), with an output of 1.

7. What are the assumptions of logistic regression?

a) No Multicollinearity (Before training model)

The independent variables must be unrelated to one another. That is, there should be minimal or no multicollinearity in the model.

b) Linearity (After training model)

Independent variables should be highly correlated with log odds i.e. log(p/1-p)

8. Why is logistic regression called regression and not classification?

Reason is:

The concept behind Logistic regression is quite similar to Linear Regression.

This model creates a regression model to predict the likelihood that a given data entry belongs to the category 1 or 0. Logistic regression models the data using the sigmoid function, much as linear regression assumes that the data follows a linear distribution.

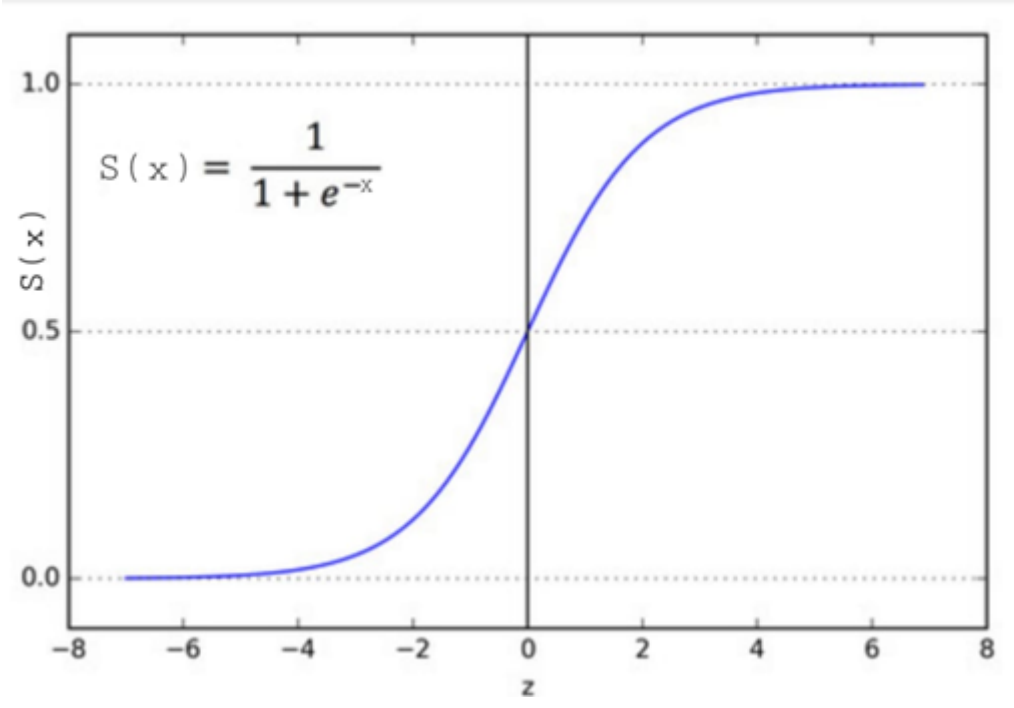
9. Explain the general intuition behind logistic regression

The intuition behind logistic regression is to transform the output of a linear regression which has a wider range, to a range that probability lies in, which is [0,1]. The transformation formula is Logit that maps a value to a number in the range [0,1].

10. Explain the significance of the sigmoid function.

Equation of sigmoid function is,

$$f(x) = \frac{1}{1 + e^{-(x)}}$$



This is the Sigmoid function, which produces an S-shaped curve. It always returns a probability value between 0 and 1. We utilize sigmoid to translate predictions to probabilities in machine learning.

To get the probabilities of dependent variable in logistic regression, we can use following function:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

11. How does Gradient Descent work in Logistic Regression?

Similar to Linear Regression, we define a cost function that estimates the deviation between the model's prediction and the original target and minimise it using gradient descent by updating the original m and c.

This ensures that we can use these m and c to make future classifications using the model. The continuous output is converted to a probabilistic output using the sigmoid function.

```
z = mx + c

p(z) = 1 / (1 + e^-z)

LL = -y * log(p(z)) - (1-y) * log(1-p(z))
```

Gradient descent algorithm calculate d/dm (LL) and d/dc (LL) and tries to minimize the cost function

```
mnew = mold - L * d/dm (LL)
cnew = cold - L * d/dc (LL)
```

12. What are outliers and how can the sigmoid function mitigate the problem of outliers in logistic regression?

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

Sigmoid function always return values between 0 and 1, no matter what the input to it is. So even if we have outliers, we will get its corresponding output value and sigmoid will covert this value which will lie in the range 0 to 1.

13. What are the outputs of the logistic model and the logistic function?

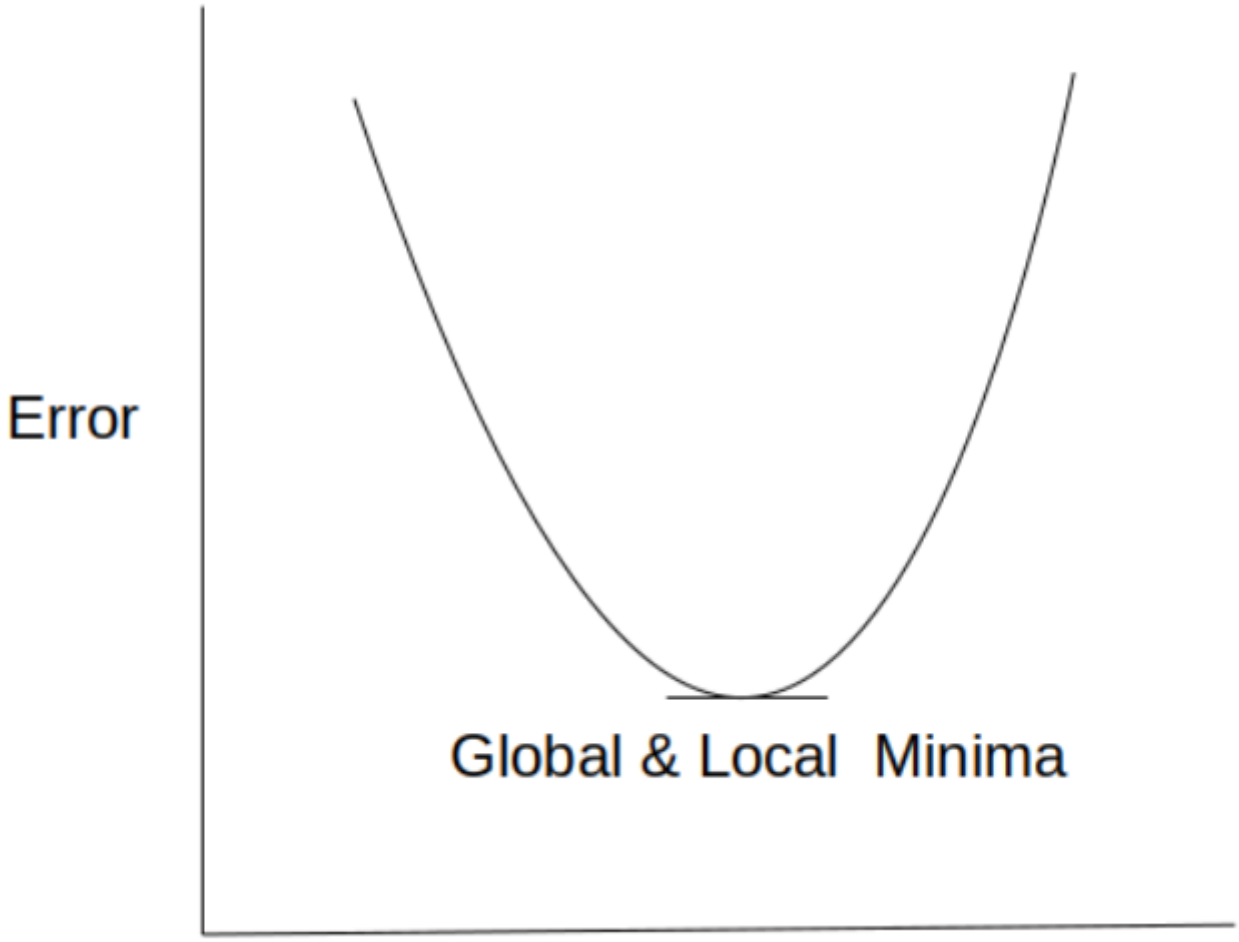
```
logistic model log(p/(1-p)) = m1x1 + m2x2 + m3x3 +..... + mnxn + c
```

Logistical model is represented as log odds. log odds is ratio of chances of success to the chances of failures.

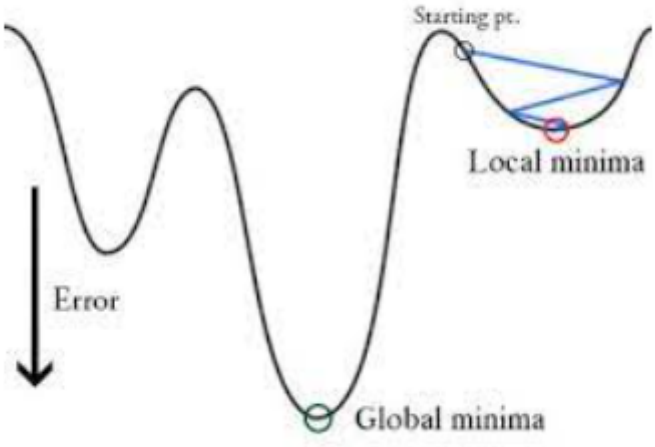
```
logistic function :  1 / (1 + e^-( m1x1 + m2x2 + m3x3 +..... + mnxn + c))
```

14. Why can’t we use Mean Square Error (MSE) as a cost function for logistic regression?

In linear regression, we used MSE as a cost function. By using gradient descent algorithm we find gradients of cost function w.r.t. m and c and finally we get best values of m and c. Since we have a convex graph now we don’t need to worry about local minima. A convex curve will always have only 1 minima.



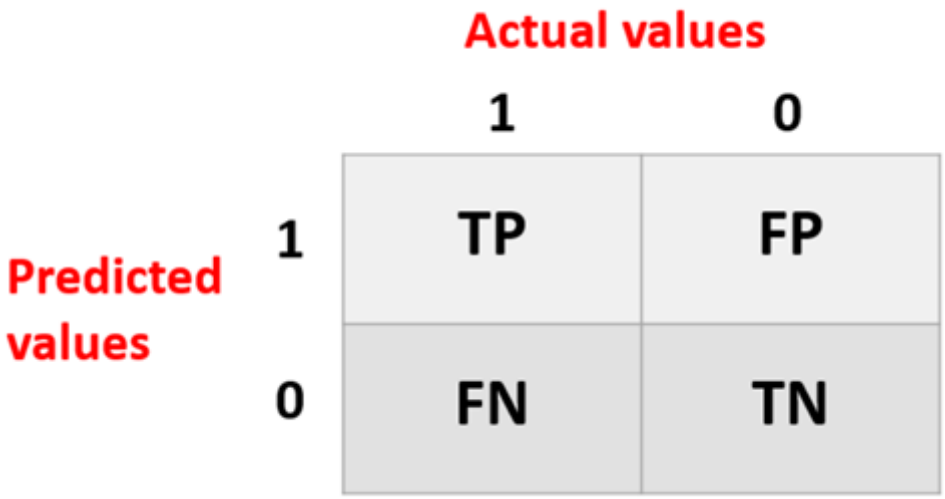
n logistic regression Yi is a non-linear function (Ŷ = 1/1+ e-z). If we use this in the above MSE equation then it will give a non-convex graph with many local minima as shown



15. What is the Confusion Matrix?

Confusion Matrix is the visual representation of the Actual VS Predicted values. It measures the performance of our Machine Learning classification model and looks like a table-like structure.

This is how a Confusion Matrix of a binary classification problem looks like :



It gives us the exact idea of how much the data from actual positive and negative class is corrrectly/incorrectly classified.

16. How do you define a classification report?

The classification report is used to measure the quality of predictions from a classification algorithm and it displays the precision, recall, F1, and support scores for the model.

17. What are the false positives and false negatives?

FP : False Positive : The values which were actually negative but falsely predicted as positive.

FN : False Negative: The values which were actually positive but falsely predicted as negative.

18. What are the true positive rate (TPR) and false-positive rate (FPR)?

TPR : Proportion of positive class got correctly classified by a classifier

FPR : Proportion of negative class got incorrectly classified by a classifier

19. What is the false-positive rate (FPR) and false-negative rate (FNR)?

FPR : Proportion of negative class got incorrectly classified by a classifier

FNR : Proportion of Positive class got incorrectly classified by a classifier

20. What are precision and recall? Explain the importance with examples?

precision : Precision checks how many outcomes are actually positive outcomes out of the total positively predicted outcomes

Recall : Recall is a measure to check correctly positive predicted outcomes out of the total number of positive outcomes.

21. What is the purpose of the precision-recall curve?

Much like the ROC curve, The precision-recall curve is used for evaluating the performance of binary classification algorithms. It is often used in situations where classes are heavily imbalanced. Also like ROC curves, precision-recall curves provide a graphical representation of a classifier's performance across many thresholds, rather than a single value

22. What is the f1 score and Explain its importance?

The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

F1 = 2 * (precision * recall) / (precision + recall)

In the multi-class and multi-label case, this is the average of the F1 score of each class with weighting depending on the average parameter.
F1-score is a machine learning model performance metric that gives equal weight to both the Precision and Recall for measuring its performance in terms of accuracy.

23. Write the equation and calculate the precision and recall rate.

```
precision = TP / (TP + FP)
Recall = TP / (TP + FN)

where,
    TP : True Positive :  The values which were actually positive and correctly predicted as positive.

    FP : False Positive : The values which were actually negative but falsely predicted as positive.

    FN : False Negative: The values which were actually positive but falsely predicted as negative.
```

24. How can you calculate accuracy using a confusion matrix?

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

where,
    TP : True Positive :  The values which were actually positive and correctly predicted as positive.

    FP : False Positive : The values which were actually negative but falsely predicted as positive.

    FN : False Negative: The values which were actually positive but falsely predicted as negative.

    TN : True Negative :  The values which were actually negative and correctly predicted as negative.
```

25. What is sensitivity?

```
    Sensitivity is a measure to check correctly positive predicted outcomes out of the total number of positive outcomes.
    Sensitivity or recall is important when you are concerned with identifying positive outcomes and the cost of a false positive is low – meaning we are fine picking up some false positives as long as we identify as many actual positives as possible. If we are predicting whether a patient has cancer or not, it is important that the sensitivity be incredibly high so that we can capture as many positive cases as possible, even if it means we pull in a few patients who don’t actually have cancer. The cost of a false positive here is low.
```

26. What is Specificity?

Specificity is the ratio of true negatives to all negative outcomes. This metric is of interest if you are concerned about the accuracy of your negative rate. Consider the example of a medical test for diagnosing a disease. Specificity relates to the test's ability to correctly reject healthy patients without a condition. Specificity of a test is the proportion of those who truly do not have the condition who test negative for the condition.

sensitivity = TN / (TN + FP)

27. What is ROC Curve?

ROC or Receiver Operating Characteristic curve represents a probability graph to show the performance of a classification model at different threshold levels. In the curve, TPR is plotted on Y-axis, whereas FPR is on the X-axis.

28. What is the importance of the ROC curve?

ROC curve helps us select the value of threshold as per the requirement.

If we have a problem statement from healthcare domain, where we expect our model to have high sensitivity (means our focus will be model should have more correct classification for actual positive class than incorrect classification), we should select a threshold value with highest tpr and comparatively low fpr.

If we have problem statement like to classify mail as spam or not spam, there we will expect actual not spam mail should not be incorrectly classified as spam, because we may miss important mail. In such a scenario, we consider precision as a evaluation metrics. So, value of False positive should be less. Here, we will select lowest fpr and comparatively high tpr.

29. What are the advantages of ROC Curve?

a. Helpful in determining optimum threshold value
b. Used to get Area Under Curve (AUC)
c. Used to plot ROC for multiple algorithms, by using AUC we can make decision on selection of algorithm.

30. What is Overfitting?

When a model performs very well for training data but has poor performance with test data (new data), it is known as overfitting. In this case, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data. If model is Overfitting, it has low bias and high variance.

31. What is Underfitting?

When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions. Underfitting model has high bias and low variance.

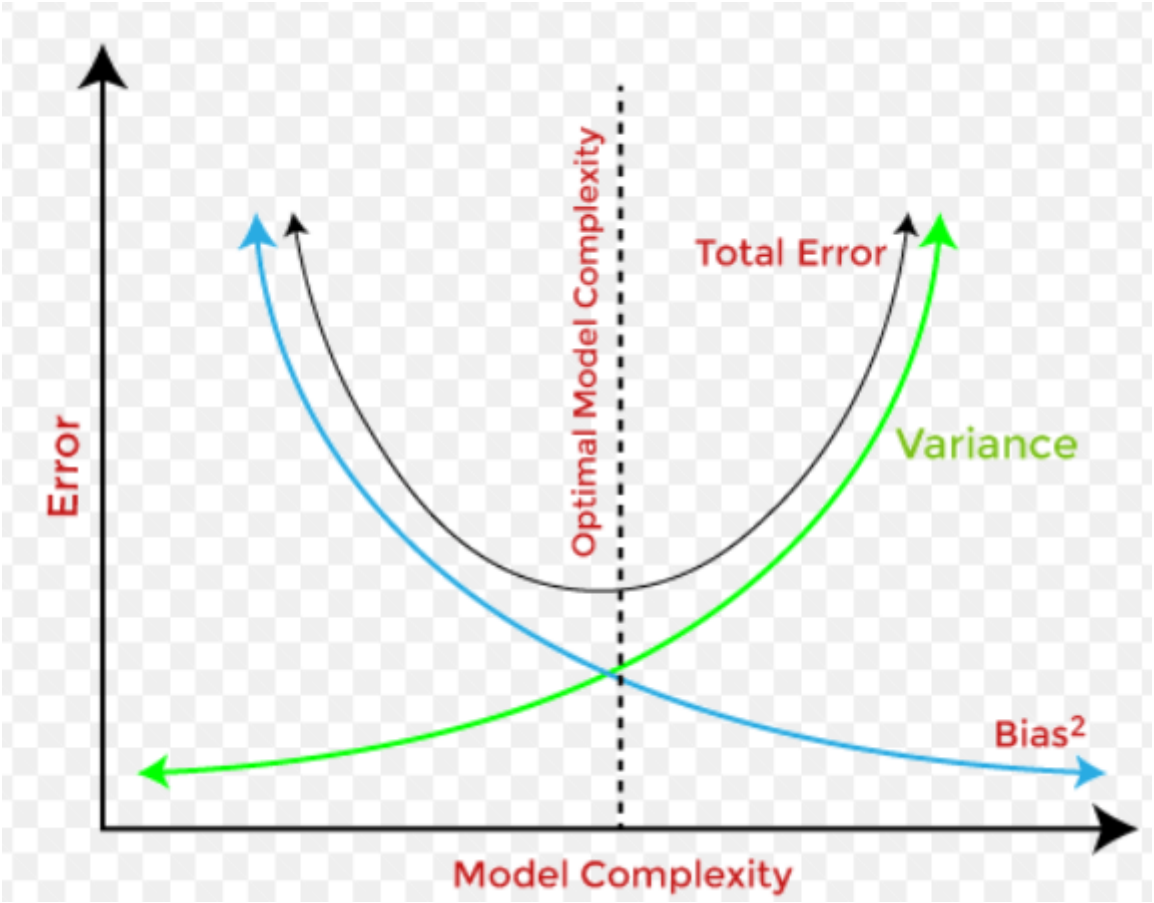
32. What is Bias and Variance in Machine Learning?

Bias can be referred the difference between our actual and predicted values. Bias is the simple assumptions that our model makes about our data to be able to predict new data.
When the Bias is high, assumptions made by our model are too basic, the model can't capture the important features of our data. This means that our model hasn't captured patterns in the training data and hence cannot perform well on the testing data too.

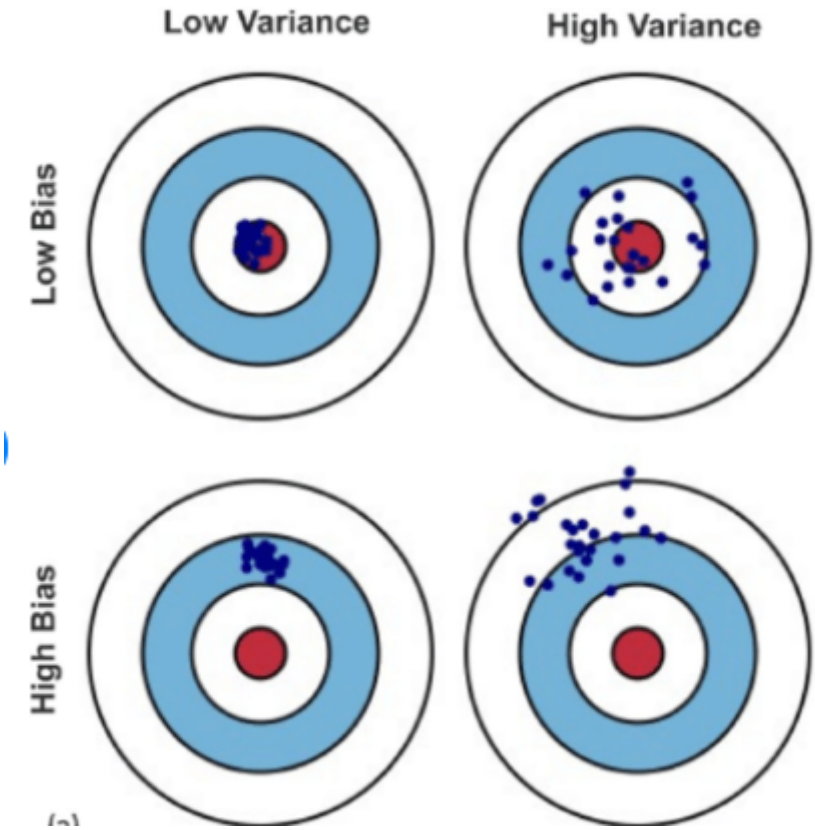
variance as the model's sensitivity to fluctuations in the data. Our model may learn from noise. Our model will perform really well on testing data and get high accuracy but will fail to perform on new, unseen data. New data may not have the exact same features and the model won't be able to predict it very well.

33. What is the tradeoff between Bias and Variance?

For any model, we have to find the perfect balance between Bias and Variance. This just ensures that we capture the essential patterns in our model while ignoring the noise present it in. This is called Bias-Variance Tradeoff. It helps optimize the error in our model and keeps it as low as possible.



34. Explain Bias and variance using the bulls-eye diagram



The above bull's eye graph helps explain bias and variance tradeoff. The best fit is when the data is concentrated in the center, i.e. at the bull's eye. We can see that as we get farther and farther away from the center, the error increases in our model. The best model is one where bias and variance are both low.

35. What is L1 and L2 regularization?

If a regression model uses the L1 Regularization technique, then it is called Lasso Regression. L1 regularization adds a penalty that is equal to the absolute value of the magnitude of the coefficient. This regularization type can result in sparse models with few coefficients. Some coefficients might become zero and get eliminated from the model. Larger penalties result in coefficient values that are closer to zero (ideal for producing simpler models).
This type of regression is used when the dataset shows high multicollinearity or when you want to automate variable elimination and feature selection.

Mathematical equation of Lasso Regression :
Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients)

Where,

λ denotes the amount of shrinkage.
The bias increases with increase in λ
variance increases with decrease in λ

Ridge regression adds “squared magnitude” of coefficient as penalty term to the loss function. Here the highlighted part represents L2 regularization element.

Mathematical equation of Lasso Regression :
Residual Sum of Squares + λ * (Sum of the square of the magnitude of coefficients)

36. How do you find the best alpha for ridge regression?

By using hyperparameter tuning (GridSearchCV and RandomizedSearchCV)

37. Does scaling affect logistic regression?

It is not mandatory to do feature scaling in logistic regression. But if we don t use it, then the difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale will help the gradient descent converge more quickly towards the minima.