Assignment- 16 Logistic Regression

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

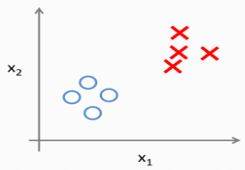
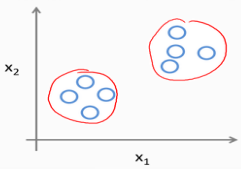
Data Science

Data Science is a process of extracting meaningful information from data using scientific methods.

Scientific Methods :-

- Machine Learning
- Deep Learning
- Natural language processing
- Statistics
- Generating AI
- Visualization

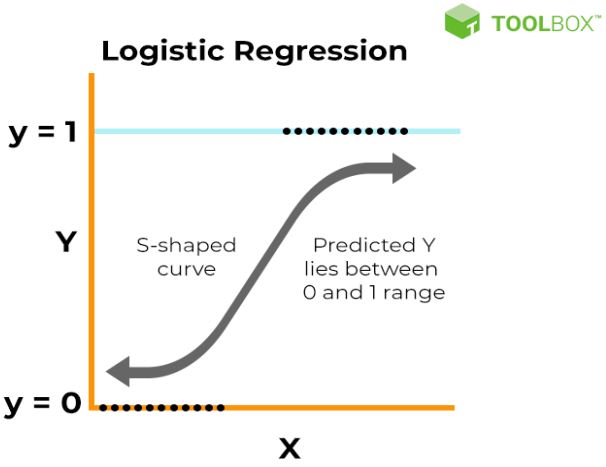
Difference between Supervised ML and Unsupervised ML

Supervised	Unsupervised
Input Data is labelled	Input Data is Unlabelled
Uses training Dataset	Uses just input dataset
Data is classified based on training dataset	Uses properties of given data to classify it.
Used for prediction	Used for Analysis
Divided into two types Regression & Classification	Divided into two types Clustering & Association
Known number of classes	Unknown number of classes
	
Use off-line analysis of data	Use Real-Time analysis of data

2. **What is logistic regression?**

- It is a supervised machine learning classification algorithm which is used to find probabilities of dependent variable.

- logistic regression is a linear model.
- Logistic regression is a probabilistic model.
- It is a Binary Classifier
- Logistic regression is employed when the dependent variable is binary or categorical.
- The output is transformed using the logistic function to ensure prediction lie between 0 to 1.



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

-Multiclass classification involves three or more classes or categories. These classes represent distinct outcomes.
- To deal with Multiclass classification we use One-vs-Rest(ovr) approach.

One-vs- Rest(ovr)

- In this approach, need create separate binary logistic regression models for each class, treating one class as the positive class and the rest of the classes as the negative class. This results in a binary classification problem for each class. During prediction, need to run each instance through all the models, and the class with the highest probability is assigned as the predicted class.

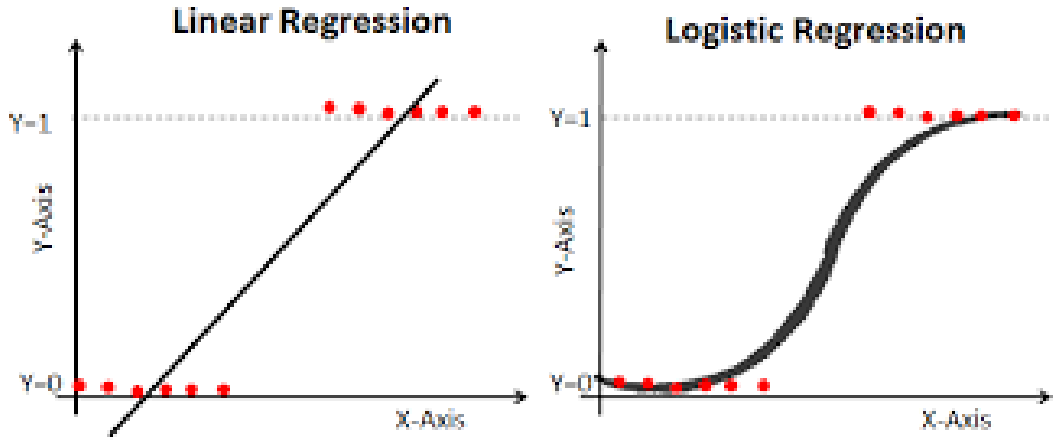
4. Why is logistic regression very popular/widely used?

- Easy to implement and interpret
- Performs well on large data and low dimensional data
- Less prone to overfitting
- Model will perform well on linearly Separable data.
- Efficient for binary classification.
- Logistic regression can handle both binary and multiclass classification problems**.**

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

- Linear regression is only used when our dependent variable have continuous data.
- If there are outliers in our dataset, then the best fit line in linear regression shifts to fit that outlier.
- Another problem with linear regression is that predicted values may be out of range, since we know probability can be between 0 to 1. If we use linear regression It may go beyond 1 & go below 0.

- To overcome these problems we use logistic regression which converts this straight best fit line in linear regression to an S-curve using sigmoid function which will always give values between 0 to 1.



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

- **Formula**

$$P(y) = 1/1+e^{-y}$$

Where,

$$Y = mx + c$$

- **Log Loss Function**

$$LL = 1/N \ (-sum(ya * log(p) + (1-ya)log(1-p))$$

Class 0 :

$LL = 1/N \ –(log(1-P))$

class 1:

$LL = 1/N \ log(P)$

7. What are the assumptions of logistic regression?

- **Linearity**

  Linear relationship between independent variable and log-odds value of probability of dependent variable.

  Log(it) = log(p / (1-p)) where p is probability of a positive outcome.

- **No Multi co-linearity**
  - No correlation between independent variables.
  - Multi co-linearity happens when two or more independent variables are highly correlated To each other, so that they do not provide the unique or individual information in regression model. If the degree of correlation Is high between variables, It can cause problems when fitting and interpreting the model.

- Multi co-linearity can be measured by Variance inflation Factor.
- If VIF = 0, then variables are not correlated.
- If VIF is between 0 to 5, then variables are moderately correlated.
- If VIF > 5, then variables are Highly correlated.

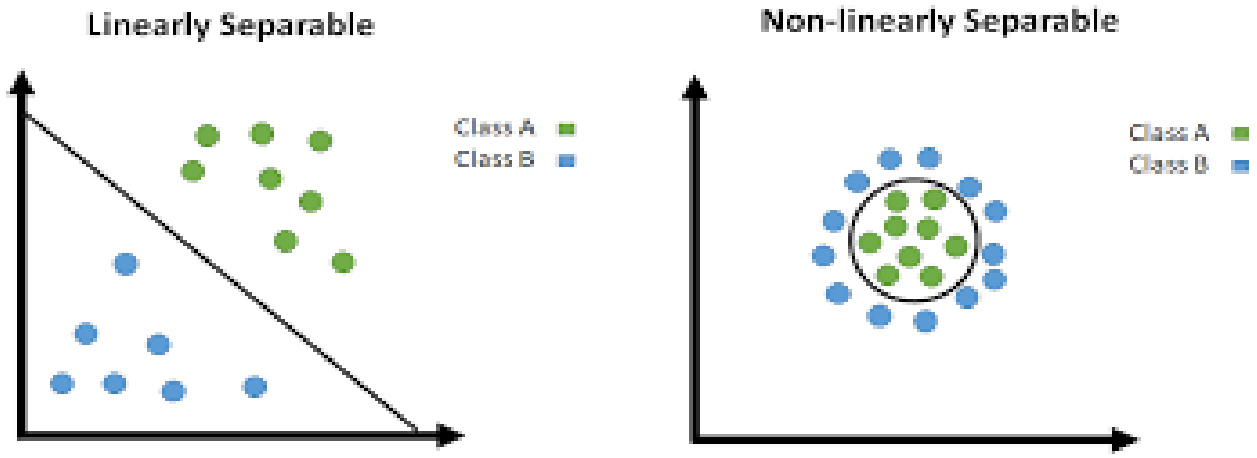
- **Large data size**
  1) Large Number of rows.
  2) Number of rows should be greater than number of columns.

- **Linearly separable Data**

  We can separate the data with Linear line according the class.

  2 Features >> Linear Line

  More than 2 features >> Hyperplane



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

Logistic regression is called logistic regression because it uses a logistic function to transform the output of the linear function into a probability value. The logistic function is a non-linear function that is shaped like an S-curve. It has a range of 0 to 1, which makes it ideal for modelling probabilities.

9. Explain the general intuition behind logistic regression.

Logistic regression predicts the probability that an instance belongs to a specific category. It uses the logistic function to transform a linear combination of input features into a probability between 0 and 1. If the probability is above a chosen threshold (like 0.5), the instance is classified as the positive class; otherwise, it's classified as the negative class. The method is named "regression" due to its historical roots but is mainly used for binary classification.

10. Explain the significance of the sigmoid function.

- Sigmoid function is a function that calculates the result value between 0 to 1 by taking the any input value.
- Sigmoid Function = **$p(y) = 1 / (1 + e^{-y})$**

  Where, p is probability of binary outcome

  $Y = mx + c$

  When y is large and positive, then probability of binary outcome is 1.0.

  When is large and negative, then probability of binary outcome is 0.0.

11. How does Gradient Descent work in Logistic Regression? Explain with its Cost Function.

- Gradient descent is an optimization algorithm which is used to find the minimum cost function, in case of logistic regression Log loss function.
- Gradient descent changes the value of m and c in such a way that it always converges tp a minimum point. Or we can say, it aims at finding the optimal values of m and c.
- It is an iterative method that finds the minimum of a function by figuring out the slopes at random point and then moving in opposite direction.
- Gradient descent used partial derivatives to find the m and c values

$$\frac{\partial}{\partial \mathrm{m}} = \frac{2}{N}\sum_{i=1}^{N} - x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \mathrm{b}} = \frac{2}{N}\sum_{i=1}^{N} - (y_i - (mx_i + b))$$

- Two things are required to find the global minimum point.
  1) Derivative – to find the direction of next step
  2) Learning rate – Magnitude of the next step.

Gradient Descent

Repeat until converge {

$$\mathrm{w} = \mathrm{w} - \alpha \left[\frac{\partial Loss}{\partial w}\right]$$

$$\mathrm{b} = \mathrm{b} - \alpha \left[\frac{\partial Loss}{\partial b}\right]$$

}

1) Start with random point.
2) Loop until convergence.
   a. Compute gradient
   b. Update
3) Return

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

Outliers:- Outliers are nothing but the odd one out meaning data points which are far away from the range of data points in dataset.

- Logistic regression model will not be much impacted due to presence of outliers. As we are using sigmoid function.
- Sigmoid function lowers the influence of outliers on the model.
- If the outliers are extreme then it may lower the model's performance, so we can handle the outliers by following ways:
  1) Either We can delete the extreme outliers.
  2) Or we can replace the outliers with mean , median and mode.
  3) Or we can also replace the outliers with upper tail and lower tail.

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

In logistic regression, the logistic model produces the estimated probabilities of an instance belonging to a certain class. These probabilities can be interpreted as the likelihood or confidence of the instance being classified as a specific class.

logistic model in logistic regression produces estimated probabilities of an instance belonging to a class

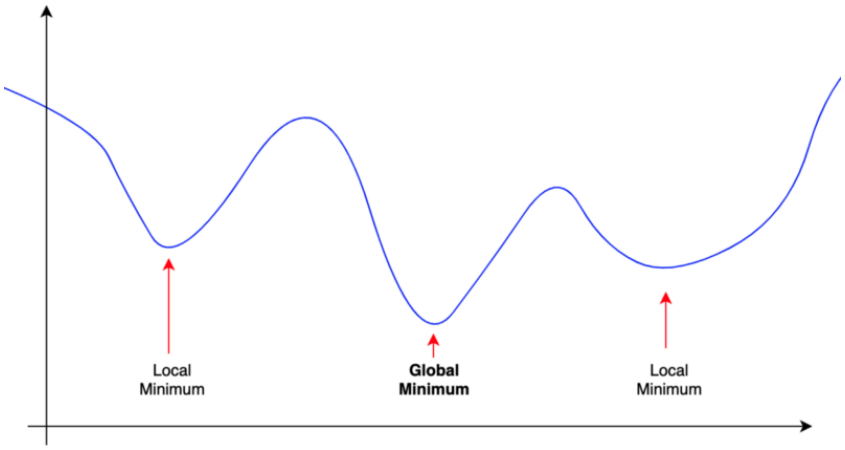
Output of Logistic Function: -

The output of the logistic function is a value between 0 and 1, indicating the probability of the instance belonging to class 1. By convention, if this probability is above a certain threshold (threshold = 0.5), the instance is classified as class 1; otherwise, it is classified as class 0.

logistic function (sigmoid function) is used to transform the linear combination of input features into these probabilities, with outputs ranging from 0 to 1.

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

- MSE is nothing but Mean Squared Error
- While using MSE we never get global minima, we always get local minima.
- MSE will always give less penalty than LogLoss function so we always prefer LogLoss function over MSE
- MSE will not give high penalty for misclassified datapoints.



15. What is the Confusion Matrix?

Confusion Matrix:

It is a visual representation of the actual vs predicted values. It measures the performance of our Machine Learning classification algorithm and looks like table like structure.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- True Positive(TP):- the values which are actually positive and predicted positive.
- False Negative(FN):- The values which are actually positive but predicted Negative.
- False Positive(FP):- the values which are actually Negative but predicted Positive.
- True Negative(TN):- The values which are actually positive and predicted Negative.

16.How do you define a classification report?

The classification report visualizer displays the precision, recall, F1, and support scores for the model.

A Classification report is used to measure the quality of predictions from a classification algorithm. The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives.

17.What are the false positives and false negatives?

- False Negative(FN) :- The values which are actually positive but predicted Negative.
- False Positive(FP) :- the values which are actually Negative but predicted Positive.

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

True Positive Rate(TPR)

It is a proportion of Positive class got correctly classified by classifier. It is also called as recall/ sensitivity.

It is ratio of True Positive values divided by total number of actual positive values.

TPR = TP/ (TP+FN)

False Positive Rate(FPR)

It is a proportion of negative class got incorrectly classified by classifier.

It is ratio of false positive values divided by total number of actual negative values.

FPR = FP/(TN+FP)

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

False Positive Rate(FPR)

It is a proportion of negative class got incorrectly classified by classifier.

It is ratio of false positive values divided by total number of actual negative values.

FPR = FP/(TN+FP)

False Negative Rate(FNR)

Proportion of positive class got incorrectly classified by the classifier called as false negative rate.

FNR = FN/(TP+FN)

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

Precision

Precision checks how many outcomes are actually positive out of total positively predicted outcomes.

Precision = TP/ (TP+FP)

Best value for precision is 1.

Importance of Precision

- Quality of Positive predictions.
  It quantifies how well the classifier performs in identifying the positive class correctly. High precision indicates a low rate of false positive predictions, meaning that when the classifier predicts an instance as positive, it is more likely to be correct.
- Decision- making and threshold selection
  Precision can aid in decision-making and threshold selection for classification tasks. By examining precision at different classification thresholds, one can choose a threshold that optimizes precision based on the specific requirements of the problem.

Example:

Let's take example of classification of email according spam or not spam, so in the case of email which are not spam but if model classified as spam, then user can lose important information, in this case false positive is important as spam email can harm the persons but if important information get lost then it will loss to person so, False positive need to improve so that user will always get the important information, in that case precision is important.

Recall

The recall is the measure of correctly positive predicted outcomes out of the total number of Positive outcomes

Recall = TP/(TP+FN)

Best value for Recall is 1

Importance of Recall

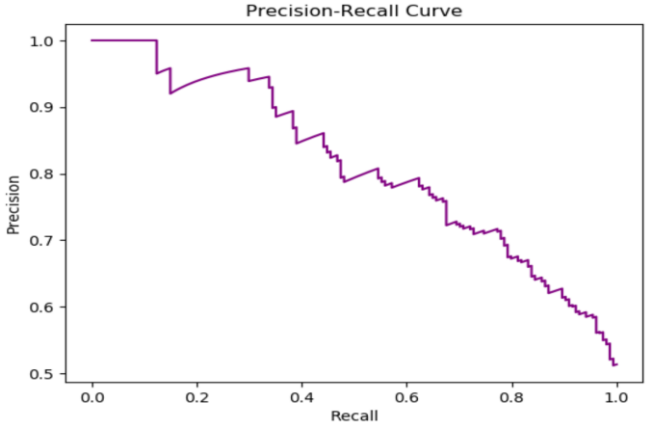
- Completeness of Positive predictions.
  Recall focuses on the ability of a classifier to capture all actual positive instances in the dataset. It measures how well the classifier avoids false negatives by correctly identifying positive instances. High recall indicates a low rate of false negatives, implying that the classifier can successfully identify a large proportion of positive instances.
- Decision making and Threshold selection
  Recall plays a role in decision-making and threshold selection for classification tasks. By examining recall at different classification thresholds, one can choose a threshold that optimizes recall based on the specific requirements of the problem.

Example:

Let us take example for cancer diseases prediction for person, it will predicted whether a person has cancer or not, if a person who have cancer but model predict a person has no cancer then person will not take treatment for cancer and may be person life gets in risk as model not predict a cancer, so it important to improve false negative values, model should able to classify positive class perfectly. In this case recall is important, a high recall means low rate of false negative values,

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

The precision-recall curve is constructed by calculating and plotting the precision against the recall for a single classifier at a variety of thresholds. For example, if we use logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class



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

F1-score

F1 score is the harmonic mean of precision and recall and it captures the contribution of both of them.

F1 Score = 2PR / P + R

Importance of f1-score

1) Balancing precision and recall
   Precision and recall are two important evaluation metrics, but they have an inverse relationship. Increasing one metric leads to a decrease in the other. The F1 score strikes a balance between precision and recall by taking their harmonic mean.
2) Model Comparison
   The F1 score enables easy comparison between different classifiers or models. When evaluating multiple models, comparing their F1 scores provides insights into their overall performance in terms of precision and recall. A higher F1 score generally indicates a better- performing model in terms of achieving a balance between precision and recall.
3) Threshold Selection
   In binary classification, the classification threshold determines whether an instance is classified as positive or negative. The F1 score can assist in selecting an appropriate threshold that optimizes the model's performance. By comparing the F1 scores at different thresholds, one can identify the threshold that maximizes the F1 score and thus achieves a good balance between precision and recall.

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

True Positive(TP) = 60
False Positive(FP) = 20
False Negatives(FN) = 30

Precision = TP/(TP+FP)
= 60/(60+20)
= 0.75

Recall = TP/(TP+FN)
= 60/(60+30)
= 0.66

Precision rate is 0.75, and Recall rate is 0.66

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

Accuracy measures the overall correctness of the classifier's prediction across both positive and negative classes. It provides an assessment of the classifier's performance in terms of the proportion of correct predictions out of all predictions made.

However, accuracy may not be the appropriate measure in some cases where data is imbalanced.

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

25.What is sensitivity?

Sensitivity is calculated by taking the ratio of TRUE POSITIVE(TP) to the sum of TRUE POSITIVE(TP) and FALSE NEGATIVES(FN), It is also called as TPR and Recall.

Sensitivity = TP/(TP+FN)

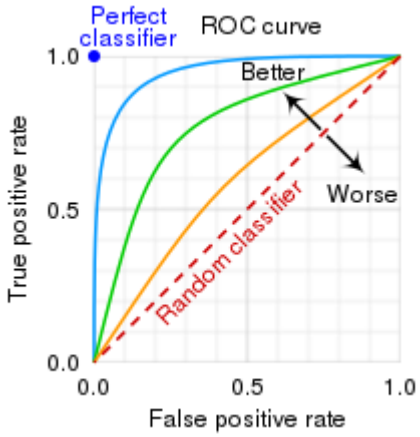
26. What is Specificity?

Specificity is calculated by taking the ratio of TRUE NEGATIVES(TN) to the sum of TRUE NEGATIVES(TN) and FALSE POSITIVES(FP).

Specificity = TN/ (TN + FP)

27. What is ROC Curve?

- The receiver operating characteristic curve is a graphical representation of trade-off between True positive rate(RECALL) anf False positive rate at various threshold values.
- Analysing the the Roc curve can help in selecting an appropriate threshold based on the desired balance.
- Changing the threshold allows us to customize the models behaviour to meet the specific needs and constraints of application.
- The higher AUC, the better the model performance at distinguishing between the positive and negative classes.



28. What is the importance of the ROC curve?

- Performance Evaluation
- Threshold Selection
- Model Comparison
- Area under the curve

29. What are the advantages of ROC Curve?

- Performance Evaluation
- Threshold Selection
- Model Comparison
- Area under the curve

30. What is Overfitting?

Over fitting :- when model is performing well only on training dataset not on testing dataset.

Training dataset Accuracy :- 98%
Testing dataset Accuracy:- 76%

OverFitting :- Low bias and High variance

31. What is Underfitting?

Underfitting:- when model is not performing well either on training dataset or on testing dataset.

Training dataset Accuracy :- 77%
Testing dataset Accuracy:- 76%

Underfitting :- High bias and low variance

32. What is Bias and Variance in Machine Learning?

Bias:-

Bias is difference between actual values and predicted values.

High bias
    Low training data accuracy
Low bias
    High training data accuracy

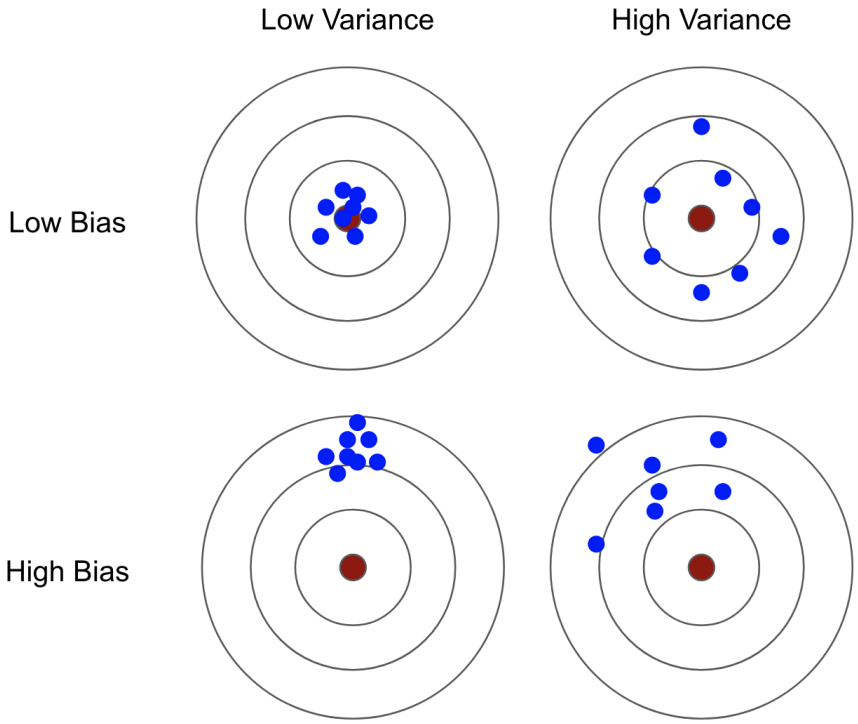
Variance:-

It is the difference between accuracy of two data sets.

High variance:-
    Difference between accuracy of training dataset and testing dataset is maximum.
Low variance:-
    Difference between accuracy of training dataset and testing dataset is minimum.

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

We can see that there is a region in the middle, where the error in both training and testing set is low and the bias and variance is in perfect balance. The above bull's eye graph helps explain bias and variance tradeoff better. The best fit is when the data is concentrated in the center, ie: at the bull's eye.



34. What is L1 and L2 regularization?

L1 Regularization:-

- Lasso regression is a type of regression that helps prevent overfitting by adding a penalty term to the traditional linear regression equation.
- This penalty encourages the model to use fewer features or set some features coefficients exactly to zero, making the model simple and more interpretable .
- It s useful when dealing with large number of features and aids in selecting the most important for predictions.

L2 Regularization:-

- Ridge regression is another type of linear regression that addresses the issue of overfitting. Similar to lasso regression adds a penalty term to traditional linear regression equation. However instead of penalizing absolute values of coefficient ridge regression penalizes the squared values of coefficients.
- It is beneficial when dealing with multi co-linearity and helps to create a more stable model.

35. How do you find the best alpha for ridge Classification?

The value of best alpha in Ridge Regression can be find through Hyperparamete Tunning.

* Default value of hyperparameter in Ridge Regression is 1.

* So with the help of cross validation, GridSearchCV and RandomSearchCV we can find best value of alpha.

- **Cross-Validation**:

This involves splitting the data into training and validation sets and evaluating the model performance on the validation set for different alpha values.
The alpha with the best performance is chosen as the best alpha.

- **Grid Search**:

This involves defining a grid of alpha values and evaluating the model performance for each alpha using cross-validation.
The alpha with the best performance, such as lowest mean squared error or highest R squared, is chosen as the best alpha.

- **RandomSearch**:

This involves defining a random values of alpha in given range and evaluating the model performance for Random alpha using cross-validation.
the alpha with best performance such as lowest mean squared error or highest R squared, from random values are choosen as best alpha

36. Does scaling affect logistic regression?

Yes, scaling can affect logistic regression. In logistic regression,
the scale of the input features can have an impact on the coefficients and the performance of the model.
It is a good practice to scale the features before fitting the logistic regression model,
and to apply the same scaling to the test data when making predictions.

37. What are the advantages and disadvantages of Logistic Regression

Advantages:-

- easy to interpret and implement
- performs well on larger data and low dimensional data.
- Less prone to overfitting.
- Model will perform well on linearly seperable data.
- Efficient for binary classification.

Disavantages

- Highly sensitive to outliers.
- Logistic regression will not perform well on small size data and high dimentional data.
- Model will not perform well on non linear datas.
- No multicolinearity.
- May overfit in presence of multiple features .