LOGISTIC REGRESSION (QUESTION AND ANSWER)

- **What is logistic regression?**
- It is a technique to analyse a data-set which has a dependent variable and one or more independent variables to predict the outcome in a binary variable, meaning it will have only two outcomes.
- The dependent variable is **categorical** in nature. Dependent variable is also referred as **target variable** and the independent variables are called the **predictors**.
- Logistic regression is a special case of linear regression where we only predict the outcome in a categorical variable. It predicts the probability of the event using the log function.
- Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

- **How many type of logistic regression?**
- four types of logistic regression analysis
- To assess the relationship between one or more predictor variable and response variable
- **Binary logistic regression:-**
- No. of categories are 2
- Characteristics are at 2 levels. E.g. - Pass or fail, Yes or No.
- **Ordinal logistic regression:-**
- No. of categorical are 3 or more
- Characteristic are at natural ordering of the levels
- E.g.- Medical Condition (critical, serious, stable, good) , survey results (disagree , neutral, agree)
- **Nominal logistic regression:-**
- No. of categories are 3 or more
- Characteristic are not as per at natural ordering of levels
- E.g.- color (red, blue, green), school subject (science, math, art)
- **Poisson logistic regression:-**
- No. of categories are 3 or more
- Characteristic are the no. of time an event occurs
- E.g. – 0, 1, 2, 3, .. etc

- **<u>There are 3 link function: -</u>**
- **Logit function**
- **Normit (probit)**
- **Gomit (complementary log-log)**
- Note: - Binary and ordinal logistic regression offer all 3 link function
- Nominal logistic regression offer only logit link function

- **Properties of Logistic Regression:**
- The dependent variable in logistic regression follows Bernoulli Distribution.
- Estimation is done through maximum likelihood.
- No R Square, Model fitness is calculated through Concordance, KS-Statistics.

- **Why cannot we use linear regression for dichotomous(binary) output?**
- The linear regression is used for unbounded output
- The linear regression does not know that the output is bounded between 0 and 1

- **What are the assumptions of logistic regression?**
- Linearity
- No endogeneity
- Normality and homoscedasticity
- No autocorrelation

LOGISTIC REGRESSION (QUESTION AND ANSWER)

- No multi-collinearity

- **What is MLE?**
- MLE stands for maximum likelihood estimate
- It is a function which estimates how likely it is that the model at hand describes the real underlying relationship of the variables.
- Bigger the MLE, the higher the probability that our model is correct.

- **Maximum Likelihood Estimation: -** The MLE is a "likelihood" maximization method . Maximizing the likelihood function determines the parameters that are most likely to produce the observed data. From a statistical point of view, MLE sets the mean and variance as parameters in determining the specific parametric values for a given model. This set of parameters can be used for predicting the data needed in a normal distribution.

- **How to measure the accuracy of logistic regression?**
- Where the prediction is < 0.5 there the predicted variable = 0. Where the prediction is >= 0.5 there the predicted variable = 1.
- Confusion matrix is used to measure the accuracy of the logistic regression.

- **What is the Sigmoid Function?**
- In order to map predicted values to probabilities, we use the sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.
- **Math :-** $S(z)=1/1+e-z$ where z = mx + c and e = base of natural log
- if the value of z goes to positive infinity then the predicted value of y will become 1 and if it goes to negative infinity then the predicted value of y will become 0. And if the outcome of the sigmoid function is more than 0.5 then we classify that label as class 1 or positive class and if it is less than 0.5 than we can classify it to negative class or label as class 0.
- We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

- These are the three fundamental concepts that you should remember next time you are using, or implementing, a logistic regression classifier:
    1. Logistic regression hypothesis
    2. Logistic regression decision boundary
    3. Logistic regression cost function
- **<u>Hypothesis Representation: -</u>**
- When using *linear regression* we used a formula of the hypothesis i.e.
- $h\Theta(x) = \beta_o + \beta_1 X$
- For logistic regression we are going to modify it a little bit i.e.
- $\sigma(Z) = \sigma(\beta_o + \beta_1 X)$
- We have expected that our hypothesis will give values between 0 and 1.
- $Z = \beta_o + \beta_1 X$
- $h\Theta(x) = \text{sigmoid}(Z)$
- i.e. $h\Theta(x) = 1/(1 + e^{-(\beta_o + \beta_1 X)})$

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

LOGISTIC REGRESSION (QUESTION AND ANSWER)

- **Decision Boundary: -** we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.
- p≥0.5,class=1
- p<0.5,class=0
- the predictions won't be perfect. This can be improved by including more features (beyond weight and height), and by potentially using a different decision boundary.
- Logistic regression decision boundaries can also be non-linear functions, such as higher degree polynomials.

- **Cost Function:-**
- We learnt about the cost function $J(\vartheta)$ in the *Linear regression*, the cost function represents optimization objective i.e. we create a cost function and minimize it so that we can develop an accurate model with minimum error.
- If we try to use the cost function of the linear regression in 'Logistic Regression' then it would be of no use as it would end up being a **non-convex** function with many local minimums, in which it would be very **difficult** to **minimize the cost value** and find the global minimum.
- For logistic regression, the Cost function is defined as:
- $-log(h\theta(x))$ *if* $y = 1$
- $-log(1-h\theta(x))$ *if* $y = 0$

- **Gradient Descent: -**
- Now the question arises, how do we reduce the cost value. Well, this can be done by using **Gradient Descent.** The main goal of Gradient descent is to **minimize the cost value.** i.e. min $J(\vartheta)$.
- Now to minimize our cost function we need to run the gradient descent function on each parameter i.e.

- **What is a Box Cox Transformation?**
- Dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

- **What are collinearity and multicollinearity?**
- Collinearity occurs when two predictor variables (e.g., x1 and x2) in a multiple regression have some correlation.
- Multicollinearity occurs when more than two predictor variables (e.g., x1, x2, and x3) are inter-correlated.

- **Regularization:-** Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting. **Regularization, significantly reduces the variance of the model, without substantial increase in its bias**.

- **Lasso vs Ridge vs Elastic net: -**
- This technique is used to prevent or reduce overfitting on sample data.
- If your linear model contains many predictor variables or if these variables are correlated, the standard OLS parameter estimates have large variance, thus making the model unreliable.
- To counter this, you can use regularization - a technique allowing to decrease this variance at the cost of introducing some bias. Finding a good bias-variance trade-off allows to minimize the model's total error.

- There are three popular regularization techniques, each of them aiming at decreasing the size of the coefficients:

LOGISTIC REGRESSION (QUESTION AND ANSWER)

- **Ridge Regression,** which penalizes sum of squared coefficients (L2 penalty).
- **penalty** means to compensation it loss and it used to shrink the value of coefficient.
- **Ridge will reduce the impact of features that are not important in predicting your y values.**
- **Ridge Formula**: Sum of Error + Sum of the squares of coefficients
- **Limitation: -** Ridge does not eliminate coefficients in your model even if the variables are irrelevant. This can be negative if you have more features than observations.

- **Lasso Regression**, which penalizes the sum of absolute values of the coefficients (L1 penalty).
- Lasso stands for Least Absolute Shrinkage Selector Operator. Lasso assigns a penalty to the coefficients in the linear model using the formula below and eliminates variables with coefficients that zero. This is called shrinkage or the process where data values are shrunk to a central point such as a mean.
- **Lasso will eliminate many features, and reduce overfitting in your linear model.**
- **Lasso Formula**: Lasso = Sum of Error + Sum of the absolute value of coefficients
- **Limitation**: - Lasso does not work well with multicollinearity.

- **Elastic Net**, a convex combination of Ridge and Lasso.
- Elastic Net combines feature elimination from Lasso and feature coefficient reduction from the Ridge model to improve your model's predictions.
- **Elastic Net Formula**: Ridge + Lasso

- Note: - C means Inverse of regularization strength

Model Validation: -

- **What do you understand by Precision and Recall?**
- Let me explain you this with an analogy:
- Imagine that, your girlfriend gave you a birthday surprise every year for the last 10 years. One day, your girlfriend asks you: 'Sweetie, do you remember all the birthday surprises from me?'
- To stay on good terms with your girlfriend, you need to recall all the 10 events from your memory. Therefore, **recall** is the ratio of the number of events you can correctly recall, to the total number of events.
- If you can recall all 10 events correctly, then, your recall ratio is 1.0 (100%) and if you can recall 7 events correctly, your recall ratio is 0.7 (70%). However, you might be wrong in some answers.

- For example, let's assume that you took 15 guesses out of which 10 were correct and 5 were wrong. This means that you can recall all events but not so precisely
- Therefore, **precision** is the ratio of a number of events you can correctly recall, to the total number of events you can recall (mix of correct and wrong recalls).
- 'From the above example (10 real events, 15 answers: 10 correct, 5 wrong), you get 100% recall but your precision is only 66.67% (10 / 15).

- **Q12. What's the difference between Type I and Type II error?**

Type I Error	Type II Error
• Type I error is a false positive. • Type I error is claiming something has happened when it hasn't.	• Type II error is a false negative. • Type II error is claiming nothing when in fact something has happened.

- **ROC Curve(*Receiver Operating Characteristic*): -** Roc curve tells us about the how good the model can distinguish between two things (*e.g If a patient has a disease or no*) and it should be used to check the

LOGISTIC REGRESSION (QUESTION AND ANSWER)

performance of an classification model.It is also called relative operating characteristic curve, because it is a comparison of two main characteristics (TPR and FPR).

- X-axis = FPR and Y-axis = TPR
- when the sensitivity increases, (1 — specificity) will also increase. This curve is known as the ROC curve.
- It is plotted between sensitivity(aka recall aka True Positive Rate) and False Positive Rate(FPR = 1-specificity).

- **AUC: -** AUC also called as **AREA UNDER CURVE**.It is used in classification analysis in order to determine which of the used models predicts the classes best. An example of its application are ROC curves.
- AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0 and if the predictions are 100% correct has an AUC of 1.
- **characteristics of AUC:**
- AUC is **scale-invariant**. It measures how well predictions are ranked, rather than their absolute values.
- AUC is **classification-threshold-invariant**. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

- **Trade-off between Sensitivity and Specificity**
- When we decrease the threshold, we get more positive values thus increasing the sensitivity. Meanwhile, this will decrease the specificity.
- Similarly, when we increase the threshold, we get more negative values thus increasing the specificity and decreasing sensitivity.
- As **Sensitivity** ↓, **Specificity** ↑
- As **Specificity** ↓, **Sensitivity** ↑

- **F-1 Score: -** It is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

- **The differences between the F1-score and the accuracy**
- Accuracy is used when the True Positives and True negatives are more important while F1-score is used when the False Negatives and False Positives are crucial
- Accuracy can be used when the class distribution is similar while F1-score is a better metric when there are imbalanced classes as in the above case.
- F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives).

- **Confusion Matrix: -**
- Confusion matrix is a tabular representation of Actual vs Predicted values.
- It is used to find out the accuracy of the model and avoid overfitting.
- Confusion matrix is one of the easiest and most intuitive metrics used for finding the accuracy of a classification model, where the output can be of two or more categories. This is the most popular method used to evaluate logistic regression.

LOGISTIC REGRESSION (QUESTION AND ANSWER)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

- False Positive – Type I Error
- Fasle Negative – Type II Error

- **Odds -** The odds of success are defined as the ratio of the probability of success over the probability of failure. In our example, the odds of success are .8/.2 = 4. That is to say that the odds of success are 4 to 1. If the probability of success is .5, i.e., 50-50 percent chance, then the odds of success is 1 to 1.
- **Odds range from 0 and positive infinity.
- **Log odds** - The transformation from odds to log of odds is the log transformation which is also known as logit function.
- **Sigmoid function** - It is inverse of logit function
- We can convert converse log of odds to find the probability by this formula