Q1. What is a logistic function? What is the range of values of a logistic function?

The logistic function is as defined below:

$$f(z)=1(1+e-z)$$

The values of a logistic function will range from 0 to 1. The values of Z will vary from $-\infty$ to $+\infty$.

Q2. 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 $-\infty$ to $+\infty$ 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.

Q3. What is the formula for the logistic regression function?

In general, the formula for logistic regression is given by the following expression:

$$f(z)=1(1+e-(\beta0+\beta1X1+\beta2X2+....+\betakXk))$$

Q4. How can the probability of a logistic regression model be expressed as a conditional probability?

The conditional probability can be given as:

$$P(\text{Discrete value of target variable} | X_1, X_2, X_3 .... X_k)$$

It is the probability of the target variable to take up a discrete value (either 0 or 1 in case of binary classification problems) when the values of independent variables are given. For example, the probability an employee will attrite (target variable) given his attributes such as his age, salary, KRA's, etc.

Q5. What are odds?

It is the ratio of the probability of an event occurring to the probability of the event not occurring. For example, let's assume that the probability of winning a lottery is 0.01. Then, the probability of not winning is 1 - 0.01 = 0.99.

Now, as per the definition,

The odds of winning the lottery = (Probability of winning)/(Probability of not winning)

The odds of winning the lottery = 0.01/0.99

Hence, the odds of winning the lottery is 1 to 99, and the odds of not winning the lottery is 99 to 1.

Q6. Why can't linear regression be used in place of logistic regression for binary classification?

The reasons why linear regressions cannot be used in case of binary classification are as follows:

Distribution of error terms: The distribution of data in the case of linear and logistic regression is different. Linear regression assumes that error terms are normally distributed. In the case of binary classification, this assumption does not hold true.

Model output: In linear regression, the output is continuous. In the case of binary classification, an output of a continuous value does not make sense. For binary classification problems, linear regression may predict values that can go beyond 0 and 1. If we want the output in the form of probabilities, which can be mapped to two different classes, then its range should be restricted to 0 and 1. As the logistic regression model can output probabilities with logistic/sigmoid function, it is preferred over linear regression.

Variance of Residual errors: Linear regression assumes that the variance of random errors is constant. This assumption is also violated in the case of logistic regression.

Q7. What is the likelihood function?

The likelihood function is the joint probability of observing the data. For example, let's assume that a coin is tossed 100 times and you want to know the probability of getting 60 heads from the tosses. This example follows the binomial distribution formula.

p = Probability of heads from a single coin toss

n = 100 (the number of coin tosses)

x = 60 (the number of heads – success)

n - x = 40 (the number of tails)

Pr (X=60 | n = 100, p)

The likelihood function is the probability that the number of heads received is 60 in a trail of 100 coin tosses, where the probability of heads received in each coin toss is p. Here the coin toss result follows a binomial distribution.

This can be reframed as follows:

Pr(X=60|n=100, p) = $c \times p_{60} \times (1-p)_{100-60}$

c = constant

p = unknown parameter

The likelihood function gives the probability of observing the results using unknown parameters.

Q8. What are the outputs of the logistic model and the logistic function?

The logistic model outputs the logits, i.e. log odds; and the logistic function outputs the probabilities.

$$\text{Logistic model} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

The output of the same will be logits.

$$\text{Logistic function} = f(z) = 1(1+e^{-(\beta 0+\beta 1X1+\beta 2X2+\beta 3X3+\ldots+\beta nXn)})$$

The output, in this case, will be the probabilities.

Q9. How to interpret the results of a logistic regression model? Or, what are the meanings of the different betas in a logistic regression model?

$\beta 0$ is the baseline in a logistic regression model. It is the log odds for an instance when all the attributes $(X1,X2,X3,...,Xn)$ are zero. In practical scenarios, the probability of all the attributes being zero is very low. In another interpretation, $\beta 0$ is the log odds for an instance when none of the attributes is taken into consideration.

All the other Betas are the values by which the log odds change by a unit change in a particular attribute by keeping all other attributes fixed or unchanged (control variables).

Q10. What is odds ratio?

Odds ratio is the ratio of odds between two groups. For example, let's assume that you are trying to ascertain the effectiveness of a medicine. You administered this medicine to the 'intervention' group and a placebo to the 'control' group.

$$\text{Odds Ratio (OR)} = \frac{\text{Odds of the Intervention Group}}{\text{Odds of the Control Group}}$$

Interpretation

If odds ratio = 1, then there is no difference between the intervention group and the control group.

If the odds ratio is greater than 1, then the control group is better than the intervention group.

If the odds ratio is less than 1, then the intervention group is better than the control group.

Q11. What is the formula for calculating odds ratio?

The formula can be given as:

$$OR_{X1,X0} = e^{\sum_{ki=1} \beta i (X1i - X0i)}$$

In the formula above, X_1 and X_0 stand for two different groups for which the odds ratio needs to be calculated. X_{1i} stands for the instance '*i*' in group X_1. X_{oi} stands for the instance '*i*' in group X_0.stands for the coefficient of the logistic regression model. Note that the baseline is not included in this formula.

Q1. What is the Maximum Likelihood Estimator (MLE)?

The MLE chooses those sets of unknown parameters (estimator) that maximise the likelihood function. The method to find the MLE is to use calculus and setting the derivative of the logistic function with respect to an unknown parameter to zero, and solving it will give the MLE. For a binomial model, this will be easy, but for a logistic model, the calculations are complex. Computer programs are used for deriving MLE for logistic models.

(Here's another approach to answering the question.)

MLE is a statistical approach to estimate the parameters of a mathematical model. MLE and ordinary square estimation give the same results for linear regression if the dependent variable is assumed to be normally distributed. MLE does not assume anything about independent variables.

Q2. What are the different methods of MLE and when is each method preferred?

In the case of logistic regression, there are two approaches to MLE. They are conditional and unconditional methods. Conditional and unconditional methods are algorithms that use different likelihood functions. The unconditional formula employs the joint probability of positives (for example, churn) and negatives (for example, non-churn). The conditional formula is the ratio of the probability of observed data to the probability of all possible configurations.

The unconditional method is preferred if the number of parameters is lower compared to the number of instances. If the number of parameters is high compared to the number of

instances, then conditional MLE is to be preferred. Statisticians suggest that conditional MLE is to be used when in doubt. Conditional MLE will always provide unbiased results.

Q3. What are the advantages and disadvantages of conditional and unconditional methods of MLE?

Conditional methods do not estimate unwanted parameters. Unconditional methods estimate the values of unwanted parameters also. Unconditional formulas can directly be developed with joint probabilities. This cannot be done with conditional probability. If the number of parameters is high relative to the number of instances, then the unconditional method will give biased results. Conditional results will be unbiased in such cases.

Q4. What is the output of a standard MLE program?

The output of a standard MLE program is as follows:

Maximised likelihood value: This is the numerical value obtained by replacing the unknown parameter values in the likelihood function with the MLE parameter estimator.

Estimated variance-covariance matrix: The diagonal of this matrix consists of the estimated variances of the ML estimates. The off-diagonal consists of the covariances of the pairs of the ML estimates.

Q5. Why can’t we use Mean Square Error (MSE) as a cost function for logistic regression?

In logistic regression, we use the sigmoid function and perform a non-linear transformation to obtain the probabilities. Squaring this non-linear transformation will lead to non-convexity with local minimums. Finding the global minimum in such cases using gradient descent is not possible. Due to this reason, MSE is not suitable for logistic regression. Cross-entropy or log loss is used as a cost function for logistic regression. In the cost function for logistic regression, the confident wrong predictions are penalised heavily. The confident right predictions are rewarded less. By optimising this cost function, convergence is achieved.

Q1. What is accuracy?

Accuracy is the number of correct predictions out of all predictions made.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Predictions}}$$

Q2. Why is accuracy not a good measure for classification problems?

Accuracy is not a good measure for classification problems because it gives equal importance to both false positives and false negatives. However, this may not be the case in most business problems. For example, in the case of cancer prediction, declaring cancer as benign is more serious than wrongly informing the patient that he is suffering from cancer. Accuracy gives equal importance to both cases and cannot differentiate between them.

Q3. What is the importance of a baseline in a classification problem?

Most classification problems deal with imbalanced datasets. Examples include telecom churn, employee attrition, cancer prediction, fraud detection, online advertisement targeting, and so on. In all these problems, the number of the positive classes will be very low when compared

to the negative classes. In some cases, it is common to have positive classes that are less than 1% of the total sample. In such cases, an accuracy of 99% may sound very good but, in reality, it may not be.

Here, the negatives are 99%, and hence, the baseline will remain the same. If the algorithms predict all the instances as negative, then also the accuracy will be 99%. In this case, all the positives will be predicted wrongly, which is very important for any business. Even though all the positives are predicted wrongly, an accuracy of 99% is achieved. So, the baseline is very important, and the algorithm needs to be evaluated relative to the baseline.

Q4. What are false positives and false negatives?

False positives are those cases in which the negatives are wrongly predicted as positives. For example, predicting that a customer will churn when, in fact, he is not churning.

False negatives are those cases in which the positives are wrongly predicted as negatives. For example, predicting that a customer will not churn when, in fact, he churns.

Q5. What are the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR)?

TPR refers to the ratio of positives correctly predicted from all the true labels. In simple words, it is the frequency of correctly predicted true labels.

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

TNR refers to the ratio of negatives correctly predicted from all the false labels. It is the frequency of correctly predicted false labels.

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

FPR refers to the ratio of positives incorrectly predicted from all the true labels. It is the frequency of incorrectly predicted false labels.

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

FNR refers to the ratio of negatives incorrectly predicted from all the false labels. It is the frequency of incorrectly predicted true labels.

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

Q6. What are sensitivity and specificity?

Specificity is the same as true negative rate, or it is equal to 1 – false positive rate. It tells you out of all the actual '0' labels, how many were correctly predicted.

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

Sensitivity is the true positive rate. It tells you out of all the actual '1' labels, how many were correctly predicted.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Q7. What are precision and recall?

Precision is the proportion of true positives out of predicted positives. To put it in another way, it is the accuracy of the prediction. It is also known as the 'positive predictive value'.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall is the same as the true positive rate (TPR) or the sensitivity.

$$\text{Recall} = \frac{TP}{TP+FN}$$

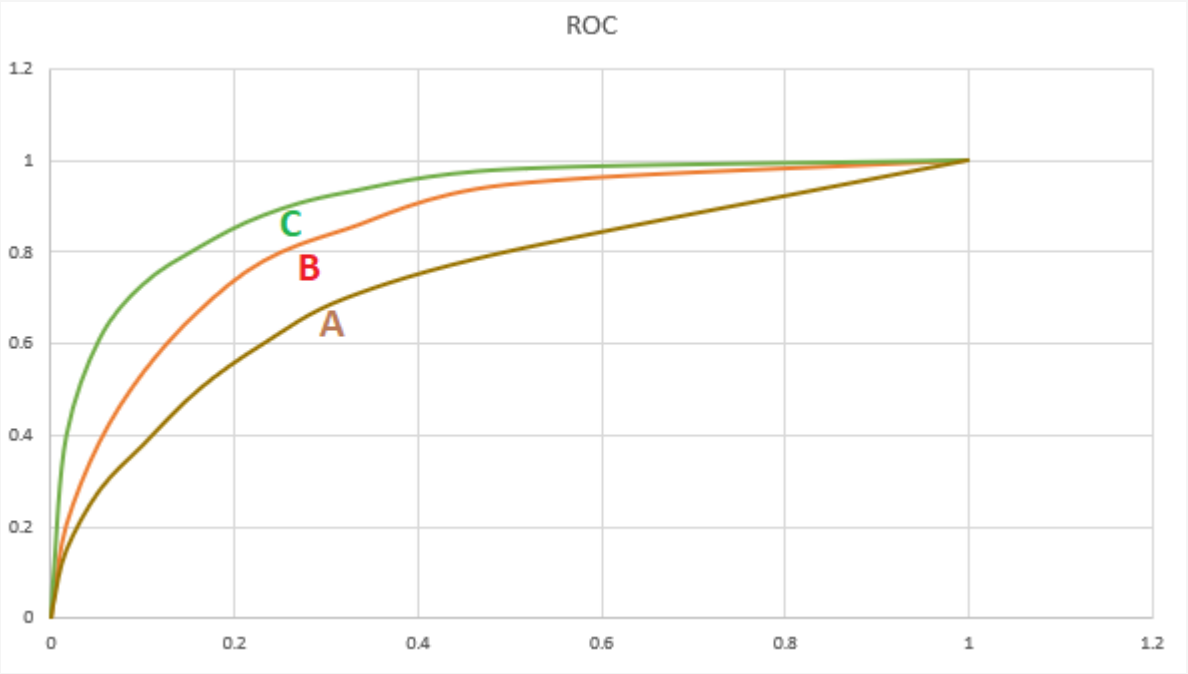
Q8. What is F-measure?

It is the harmonic mean of precision and recall. In some cases, there will be a trade-off between the precision and the recall. In such cases, the F-measure will drop. It will be high when both the precision and the recall are high. Depending on the business case at hand and the goal of data analytics, an appropriate metric should be selected.

$$F\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Q9. Explain the use of ROC curves and the AUC of an ROC Curve.

An ROC (Receiver Operating Characteristic) curve illustrates the performance of a binary classification model. It is basically a TPR versus FPR (true positive rate versus false positive rate) curve for all the threshold values ranging from 0 to 1. In an ROC curve, each point in the ROC space will be associated with a different confusion matrix. A diagonal line from the bottom-left to the top-right on the ROC graph represents random guessing. The Area Under the Curve (AUC) signifies how good the classifier model is. If the value for AUC is high (near 1), then the model is working satisfactorily, whereas if the value is low (around 0.5), then the model is not working properly and just guessing randomly. From the image below, curve C (green) is the best ROC curve among the three and curve A (brown) is the worst ROC curve among the three.



ROC Curves

Q10. How to choose a cutoff point in case of a logistic regression model?

The cutoff point depends on the business objective. Depending on the goals of your business,

the cutoff point needs to be selected. For example, let’s consider loan defaults. If the business objective is to reduce the loss, then the specificity needs to be high. If the aim is to increase the profits, then it is an entirely different matter. It may not be the case that profits will increase by avoiding giving loans to all predicted default cases. But it may be the case that the business has to disburse loans to default cases that are slightly less risky to increase the profits. In such a case, a different cutoff point, which maximises profit, will be required. In most of the instances, businesses will operate around many constraints. The cutoff point that satisfies the business objective will not be the same with and without limitations. The cutoff point needs to be selected considering all these points. If the business context doesn't matter much and you want to create a balanced model, then you use an ROC curve to see the tradeoff between sensitivity and specificity and accordingly choose an optimal cutoff point where both these values along with accuracy are decent.