# What are the key metrics used to check the performance of logistic regression?

# How do you handle missing values?

# Is logistic regression a generative or a descriptive classifier?

## What are the disadvantages of Logistic Regression?

1. How many binary classifiers would you need to implement one-vs-rest for three classes?
2. Will the decision boundary be linear or non-linear in logistic regression models?
3. In classification problems like logistic regression, classification accuracy alone is not considered a good measure. Why?
4. What is the formula for the logistic regression function?
5. Why can’t we use Mean Square Error (MSE) as a cost function for logistic regression?
6. How does logistic regression handle categorical variables?

Solutions:

1. C statistics — it represents the concordance of the model. It is the probability that an observation having an event is more than the probability that an observation having a non-event.

· Accuracy — (True positive + True negative) / Total cases

· Error Rate — (False positive + False negative) / Total cases

· Sensitivity — True positive / Total actual positive

· Specificity — True negative / Total actual negative

· Positive pred value — True positive / Total predicted positive

· Negative pred value — True negative / Total predicted negative

· KS — it measures the distance between cumulative good and cumulative bad. The maximum distance is KS.

· AUCROC — measures the performance of the model across all cut-offs. Sensitivity is on the y-axis and 1-specificty is on the x-axis

· Gain chart — positive prediction rate is on y-axis and percentage of cases allocated to event is on x-axis

1. The goal of missing value imputation is to retain all original data and score new cases

· Numerical variable — impute with mean or median and create a missing value indicator

· Categorical variable — impute with a new label

· Regression imputation — does not involve target variables and can be used when two or more variables are highly correlated. However, it may lead to over-fitting, increase computation time and increased scoring efforts

· Cluster imputation — it is condition on other variables. The cluster mean is used to replace the missing data point

1. Logistic regression is a descriptive model. Logistic [regression](https://www.projectpro.io/article/types-of-regression-analysis-in-machine-learning/410/) learns to classify by knowing what features differentiate two or more classes of objects. For example, to classify between an apple and an orange, it will learn that the orange is orange in color and an apple is not. On the other hand, a generative classifier like a Naive Bayes will store all the classes' critical features and then classify based on the features the test case best fits.
2. a) Sometimes a lot of Feature Engineering is required.

b) If the independent features are correlated with each other it may affect the performance of the classifier.

c) It is quite sensitive to noise and overfitting.

d) Logistic Regression should not be used if the number of observations is lesser than the number of features, otherwise, it may lead to overfitting.

e) By using Logistic Regression, it is tough to obtain complex relationships.

1. You would need three binary classifiers to implement one-vs-all for three classes since the number of binary classifiers is precisely equal to the number of classes with this approach. If you have three classes given by y=1, y=2, and y=3, then the three classifiers in the one-vs-all approach would consist of h(1)(x), which classifies the test cases as 1 or not 1, h(2)(x) which classifies the test cases as 2 or not 2 and so on. You can then take the results together to arrive at the correct [classification](https://www.projectpro.io/article/7-types-of-classification-algorithms-in-machine-learning/435).
2. The decision boundary is essentially a line or a plane that demarcates the boundary between the classes to which linear regression classifies the dependent variables. The shape of the decision boundary will depend entirely on the logistic regression model.

For the logistic regression model given by hypothesis function h(x)=g(Tx)where g is the sigmoid function, if the hypothesis function is h(x)=g(1+2x2+3x3)then the decision boundary is linear. Alternatively, if h(x)=g(1+2x22+3x32)then the decision boundary is non-linear.

1. Classification accuracy considers both true positives and false positives with equal significance. If this were just another machine learning problem of not too much consequence, this would be acceptable. However, when the problems involve deciding whether to consider a candidate for life-saving treatment, false positives might not be as bad as false negatives. The opposite can also be true in some cases. Therefore, while there is no single best way to evaluate a classifier, accuracy alone may not serve as a good measure.
2. f(z) = 1/(1+e-(α+1X1+2X2+….+kXk))
3. 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.
4. The inputs to a logistic regression model need to be numeric. The algorithm cannot handle categorical variables directly. So, they need to be converted into a format that is suitable for the algorithm to process. The various levels of a categorical variable will be assigned a unique numeric value known as the dummy variable. These dummy variables are handled by the logistic regression model as any other numeric value.