Logistic regression

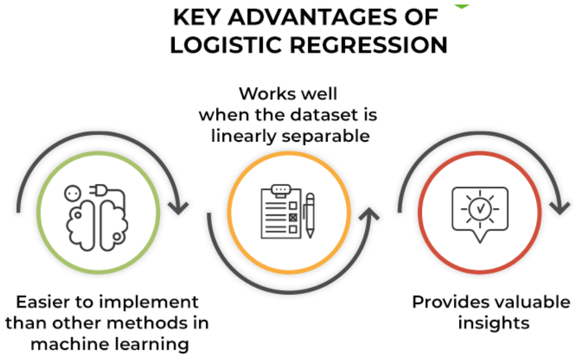
Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false
Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not
For example, 0 – represents a negative class; 1 – represents a positive class. Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1)

- Some examples of such classifications and instances where the binary response is expected or implied are:

1. Determine the probability of heart attacks: With the help of a logistic model, medical practitioners can determine the relationship between variables such as the weight, exercise, etc., of an individual and use it to predict whether the person will suffer from a heart attack or any other medical complication.

2. Possibility of enrolling into a university: Application aggregators can determine the probability of a student getting accepted to a particular university or a degree course in a college by studying the relationship between the estimator variables, such as GRE, GMAT, or TOEFL scores.

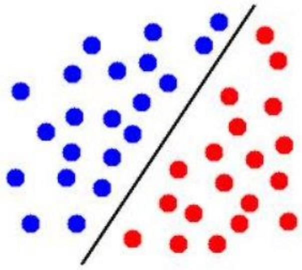
3. Identifying spam emails: Email inboxes are filtered to determine if the email communication is promotional/spam by understanding the predictor variables and applying a logistic regression algorithm to check its authenticity.


KEY ADVANTAGES OF LOGISTIC REGRESSION
Works well when the dataset is linearly separable
Easier to implement than other methods in machine learning
Provides valuable insights

Key advantages of logistic regression

The logistic regression analysis has several advantages in the field of machine learning.

1. Easier to implement machine learning methods: A machine learning model can be effectively set up with the help of training and testing. The training identifies patterns in the input data (image) and associates them with some form of output (label). Training a logistic model with a regression algorithm does not demand higher computational power. As such, logistic regression is easier to implement, interpret, and train than other ML methods.



2. Suitable for linearly separable datasets: A linearly separable dataset refers to a graph where a straight line separates the two data classes. In logistic regression, the y variable takes only two values. Hence, one can effectively classify data into two separate classes if linearly separable data is used.

3. Provides valuable insights: Logistic regression measures how relevant or appropriate an independent/predictor variable is (coefficient size) and also reveals the direction of their relationship or association (positive or negative).

Logistic Regression Equation and Assumptions

Logistic regression uses a logistic function called a sigmoid function to map predictions and their probabilities. The sigmoid function refers to an S-shaped curve that converts any real value to a range between 0 and 1.

Moreover, if the output of the sigmoid function (estimated probability) is greater than a predefined threshold on the graph, the model predicts that the instance belongs to that class. If the estimated probability is less than the predefined threshold, the model predicts that the instance does not belong to the class.

For example, if the output of the sigmoid function is above 0.5, the output is considered as 1. On the other hand, if the output is less than 0.5, the output is classified as 0. Also, if the graph goes further to the negative end, the predicted value of y will be 0 and vice versa. In other words, if the output of the sigmoid function is 0.65, it implies that there are 65% chances of the event occurring; a coin toss, for example.

The sigmoid function is referred to as an activation function for logistic regression and is defined as:

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

Equation of Logistic Regression

where,

- e = base of natural logarithms
- value = numerical value one wishes to transform

The following equation represents logistic regression:

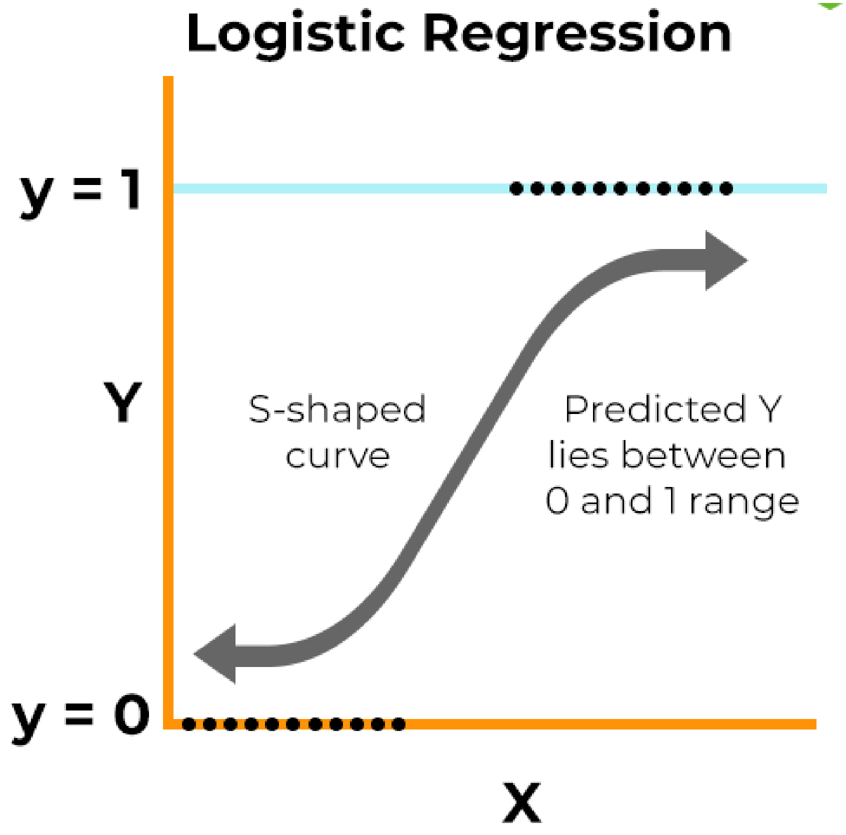
$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

Logistic Regression – Sigmoid Function

here,

- x = input value
- y = predicted output
- b0 = bias or intercept term
- b1 = coefficient for input (x)

This equation is similar to linear regression, where the input values are combined linearly to predict an output value using weights or coefficient values. However, unlike linear regression, the output value modeled here is a binary value (0 or 1) rather than a numeric value.



Key Assumptions for Implementing Logistic Regression

Key properties of the logistic regression equation

Typical properties of the logistic regression equation include:

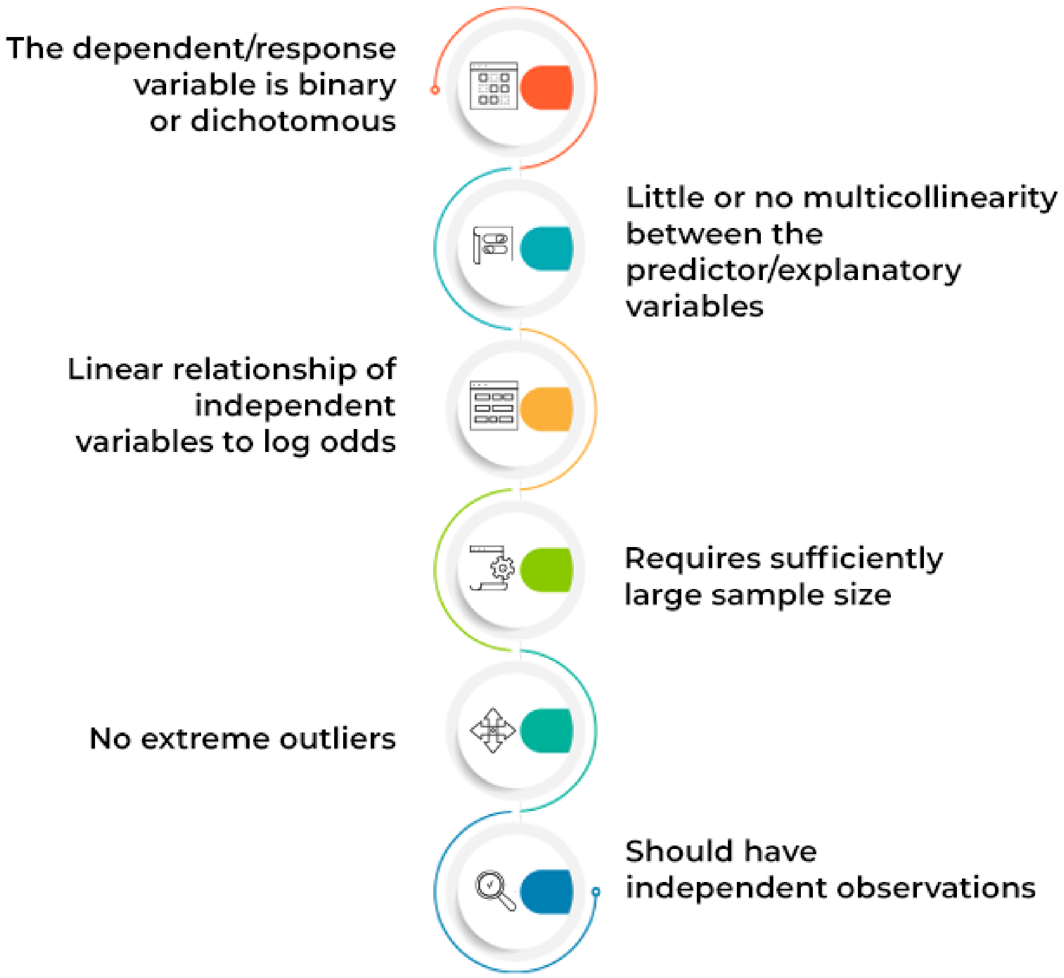
- Logistic regression's dependent variable obeys 'Bernoulli distribution'
- Estimation/prediction is based on 'maximum likelihood.'
- Logistic regression does not evaluate the coefficient of determination (or R squared) as

observed in linear regression’. Instead, the model’s fitness is assessed through a concordance.

For example, KS or Kolmogorov-Smirnov statistics look at the difference between cumulative events and cumulative non-events to determine the efficacy of models through credit scoring.

While implementing logistic regression, one needs to keep in mind the following key assumptions:

KEY ASSUMPTIONS FOR IMPLEMENTING LOGISTIC REGRESSION



The dependent/response variable is binary or dichotomous

Little or no multicollinearity between the predictor/explanatory variables

Linear relationship of independent variables to log odds

Requires sufficiently large sample size

No extreme outliers

Should have independent observations

Logistic Regression Best Practices

1. The dependent/response variable is binary or dichotomous

The first assumption of logistic regression is that response variables can only take on two possible outcomes – pass/fail, male/female, and malignant/benign.

This assumption can be checked by simply counting the unique outcomes of the dependent variable. If more than two possible outcomes surface, then one can consider that this assumption is violated.

2. Little or no multicollinearity between the predictor/explanatory variables

This assumption implies that the predictor variables (or the independent variables) should be independent of each other. Multicollinearity relates to two or more highly correlated independent variables. Such variables do not provide unique information in the regression model and lead to wrongful interpretation.

The assumption can be verified with the variance inflation factor (VIF), which determines the correlation strength between the independent variables in a regression model.

3. Linear relationship of independent variables to log odds

Log odds refer to the ways of expressing probabilities. Log odds are different from probabilities. Odds refer to the ratio of success to failure, while probability refers to the ratio of success to everything that can occur.
For example, consider that you play twelve tennis games with your friend. Here, the odds of you winning are 5 to 7 (or 5/7), while the probability of you winning is 5 to 12 (as the total games played = 12).

4. Prefers large sample size

Logistic regression analysis yields reliable, robust, and valid results when a larger sample size of the dataset is considered.
This assumption can be validated by taking into account a minimum of 10 cases considering the least frequent outcome for each estimator variable. Let's consider a case where you have three predictor variables, and the probability of the least frequent outcome is 0.30. Here, the sample size would be (10*3) / 0.30 = 100.

5. Problem with extreme outliers

Another critical assumption of logistic regression is the requirement of no extreme outliers in the dataset.
This assumption can be verified by calculating Cook's distance (Di) for each observation to identify influential

data points that may negatively affect the regression model. In situations when outliers exist, one can implement the following solutions:

- Eliminate or remove the outliers
- Consider a value of mean or median instead of outliers, or
- Keep the outliers in the model but maintain a record of them while reporting the regression results

6. Consider independent observations

This assumption states that the dataset observations should be independent of each other. The observations should not be related to each other or emerge from repeated measurements of the same individual type. The assumption can be verified by plotting residuals against time, which signifies the order of observations. The plot helps in determining the presence or absence of a random pattern. If a random pattern is present or detected, this assumption may be considered violated.

See More: 3 Ways Organizations Can Maximize ROI From AI Deployments

Types of Logistic Regression with Examples

Logistic regression is classified into binary, multinomial, and ordinal. Each type differs from the other in execution and theory. Let's understand each type in detail.

1. Binary logistic regression

Binary logistic regression predicts the relationship between the independent and binary dependent variables. Some examples of the output of this regression type may be, success/failure, 0/1, or true/false.

Examples:

1. Deciding on whether or not to offer a loan to a bank customer: Outcome = yes or no.
2. Evaluating the risk of cancer: Outcome = high or low.
3. Predicting a team's win in a football match: Outcome = yes or no.

2. Multinomial logistic regression

A categorical dependent variable has two or more discrete outcomes in a multinomial regression type. This implies that this regression type has more than two possible outcomes.

Examples:

1. Let's say you want to predict the most popular transportation type for 2040. Here, transport type equates to the dependent variable, and the possible outcomes can be electric cars, electric trains, electric buses, and electric bikes.
2. Predicting whether a student will join a college, vocational/trade school, or corporate industry.

3. Estimating the type of food consumed by pets, the outcome may be wet food, dry food, or junk food.

3. Ordinal logistic regression

Ordinal logistic regression applies when the dependent variable is in an ordered state (i.e., ordinal). The dependent variable (y) specifies an order with two or more categories or levels.

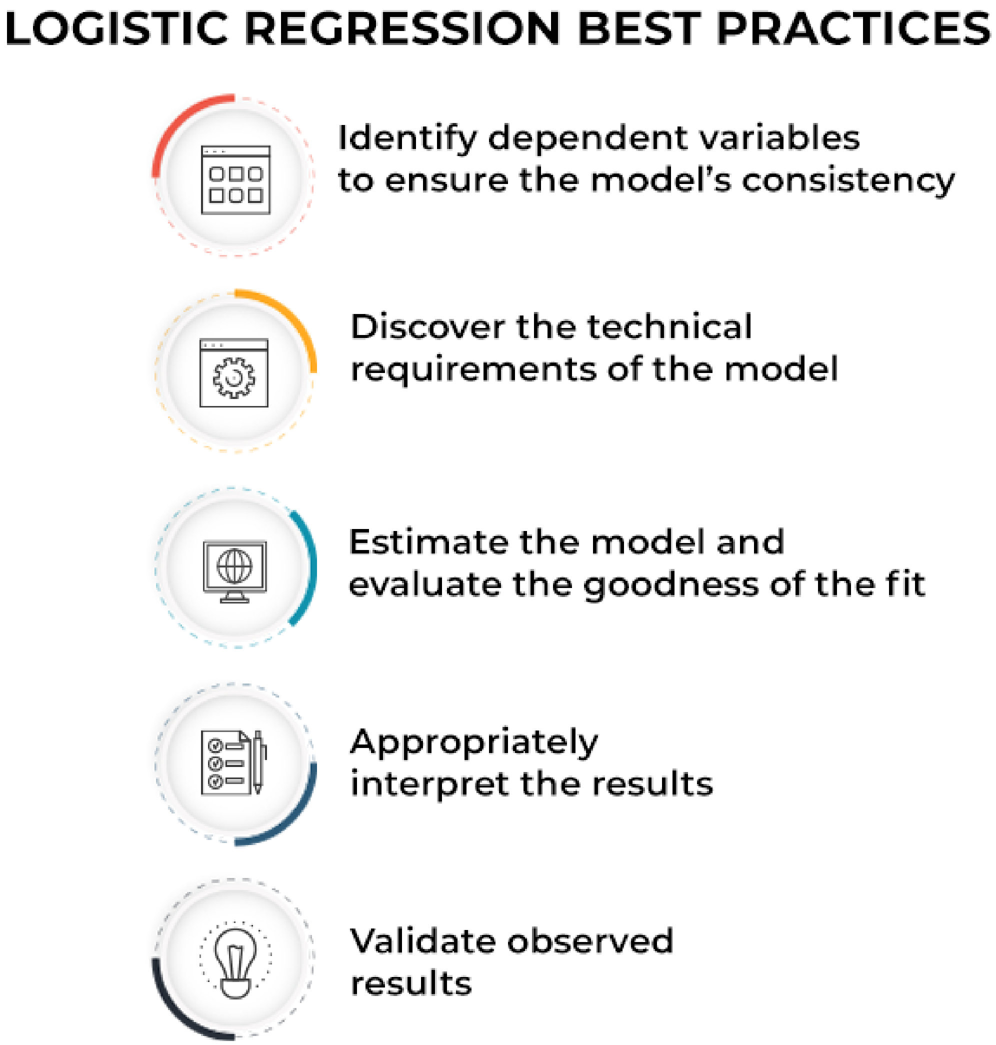
Examples: Dependent variables represent,

1. Formal shirt size: Outcomes = XS/S/M/L/XL
2. Survey answers: Outcomes = Agree/Disagree/Unsure
3. Scores on a math test: Outcomes = Poor/Average/Good

See More: **Why Machine Learning Accuracy Matters and Top Tools to Supercharge It**

Logistic Regression Best Practices

Logistic regression can produce an accurate model if some best practices are followed, from independent variable selection and choice of model building strategy to validating the model results.



LOGISTIC REGRESSION BEST PRACTICES

Identify dependent variables to ensure the model's consistency

Discover the technical requirements of the model

Estimate the model and evaluate the goodness of the fit

Appropriately interpret the results

Validate observed results

Representation of Two Logistic Regression Models

Let's understand the logistic regression best practices for 2022 in detail.

1. Identify dependent variables to ensure the model's consistency

Logistic regression performs well when one can identify a research question that reveals a naturally dichotomous dependent variable. For example, logistic regression in healthcare uses common variables such as sick/not sick, cancerous/non-cancerous, malignant/benign, and others. Medical researchers should avoid the recoding of continuous or discrete variables into dichotomous categorical variables. For example, if the variable is income per capita, recoding the income to produce two specific categories, rich versus poor, is highly inappropriate.

In technical terms, such a recoding process (converting a quantitative variable to categorical) causes a loss of information and reduces the model's estimates' consistency.

2. Discover the technical requirements of the model

Although logistic regression is a flexible statistical technique, one must keep track of the technical requirements to ensure the model's efficiency. For example, logistic regression models face problems when it comes to multicollinearity. This issue can be handled, and correlation can be minimized by considering the following strategies:

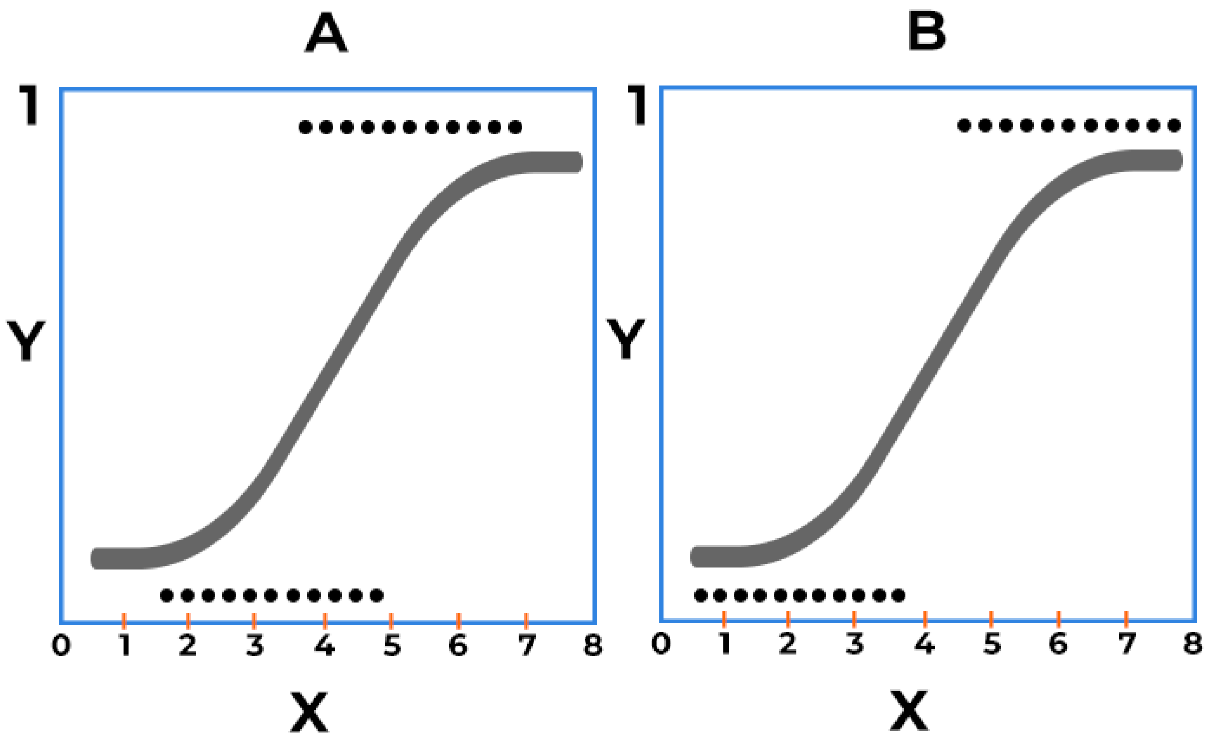
- Increase the number of observations
- Use data reduction techniques to create a synthetic measure of the original variables
- Monitor the size of samples as it is crucial in logistic regression; small samples often produce inconsistent estimates

- Exclude the extreme outliers from the model's estimation and quantify the impact of their presence on the coefficients. This ensures that atypical observations do not harm the model's fit

3. Estimate the model and evaluate the goodness of the fit

Researchers using logistic regression are also required to estimate the regression model. This involves reporting the software and sharing the replication materials, including original data, manipulated data, and computational scripts. Such practices provide transparency and make replicability of model results easier.

Upon estimating, researchers can then evaluate the fit to choose the model that excels in prediction even with minimal predictors. Not all predictors are related to the outcome. The goodness-of-fit can be tested by comparing the model having independent variables with the null model (only the intercept). Consider the figure below:



In this figure, model B represents a better fit than model A. This is because, although model A shows high variability, model B seems to be more precise.

4. Appropriately interpret the results

A logistic model is accurate when it has a fine-tuned build strategy and when the interpretation of the results produced by it is made right. Generally, a model is rated purely by analyzing the statistical significance of the estimates. However, not much attention is given to the magnitude of the coefficients. Thus, interpreting the coefficients and discussing how the results relate to the research hypothesis or question is one of the good practices for logistic regression.

Coefficients are easy to interpret in linear regression but not in logistic regression, as the estimates produced in the latter are not as intuitive. In logistic type regression, the logit transformation reveals the independent variable's impact on the variation of the dependent variable's natural logarithm of the odds.

For example, consider a coefficient of 0.4. In this case, an increase of 0.4 units is expected in the logit of y every time there's one unit increase in x. Here, it is not intuitive enough to specify that the amount in logit increased by 0.4 units with each unit increase in x. Also, it does not disclose the true relationship between the variables.

In other words, the appropriate interpretation of coefficients and the analysis of estimates is a key practice for the success of logistic regression models.

5. Validate observed results

Another critical practice that researchers can implement is validating the observed results with a subsample of the original dataset. This practice makes the model results more reliable, especially when working with smaller samples.

Result validation can help establish external validity through a separate sample or the estimation sample. External validity determines whether inferences and conclusions are valid for the model's specific population and if they can be generalized to other populations and settings. Thus, it helps represent the predicted accuracy of the designed regression model.

This approach is rarely used by professionals owing to the lack of training on the specificities of logistic regression.