Logistic Regression for Machine Learning   
[A Beginners Guide]



**Anish Mahapatra**

Contents

[Introduction 4](#_Toc113217980)

[What is Logistic Regression? 4](#_Toc113217981)

[Logistic Function (Sigmoid Function) 4](#_Toc113217982)

[Why are we discussing this for logistic regression? 5](#_Toc113217983)

[Logistic Regression Equation 6](#_Toc113217984)

[Three types of Logistic Regression 9](#_Toc113217985)

[Binary Logistic Regression 9](#_Toc113217986)

[Multinomial Logistic Regression 9](#_Toc113217987)

[Ordinal Logistic Regression 9](#_Toc113217988)

[Logistic Regression Hypothesis\*\* 10](#_Toc113217989)

[How does Logistic Regression Algorithm work? 10](#_Toc113217990)

[Sigmoid function 10](#_Toc113217991)

[When and Where to use Logistic Regression? 11](#_Toc113217992)

[Building Logistic Regression Model 11](#_Toc113217993)

[Applications of Logistic Regression 13](#_Toc113217994)

[Pros and Cons of Logistic Regression 14](#_Toc113217995)

[Pros of Logistic Regression 14](#_Toc113217996)

[Cons of Logistic Regression 15](#_Toc113217997)

[Making predictions with Logistic Regression 15](#_Toc113217998)

[Logistic Regression Assumptions 15](#_Toc113217999)

[Logistic Regression Best Practices for 2022 16](#_Toc113218000)

[Logistic vs Linear Regression 16](#_Toc113218001)

[Conclusion 17](#_Toc113218002)

[FAQs 17](#_Toc113218003)

[Q. Why is logistic regression used in Machine Learning? 17](#_Toc113218004)

[Q. Is logistic regression a model or algorithm? 17](#_Toc113218005)

[Q. Why we use logistic regression instead of linear regression? 17](#_Toc113218006)

[Q. Why logistic regression is the best model? 17](#_Toc113218007)

[Q. What algorithm is used in logistic regression? 17](#_Toc113218008)

[Author 18](#_Toc113218009)

# Introduction

A good way to start with Data Science is to understand some of the models that every Data Scientist uses. The alternative flow of thought can be to focus on the topics that interviewers ask the most in Data Science interviews. Whatever your aim is, you are in the right place. Today we will discuss one of the most requested topics in the world of Data Science!

Here at Knowledgehut, our focus is on [data science training](mailto:https://www.knowledgehut.com/data-science-courses) and [data science Bootcamp with job placement](mailto:https://www.knowledgehut.com/data-science/data-science-bootcamp-training). In this article, we will focus on understanding **logistic regression** for machine learning. Follow along to understand the logistic regression algorithm in detail. We will learn more about what is logistic regression in machine learning when to use logistic regression, how does logistic regression algorithm work and go through logistic regression from. Meanwhile, in case you are interested, you can have a look at the [knowledgehut data science bootcamp with job placement](https://www.knowledgehut.com/data-science/data-science-bootcamp-training). Let’s dive in!

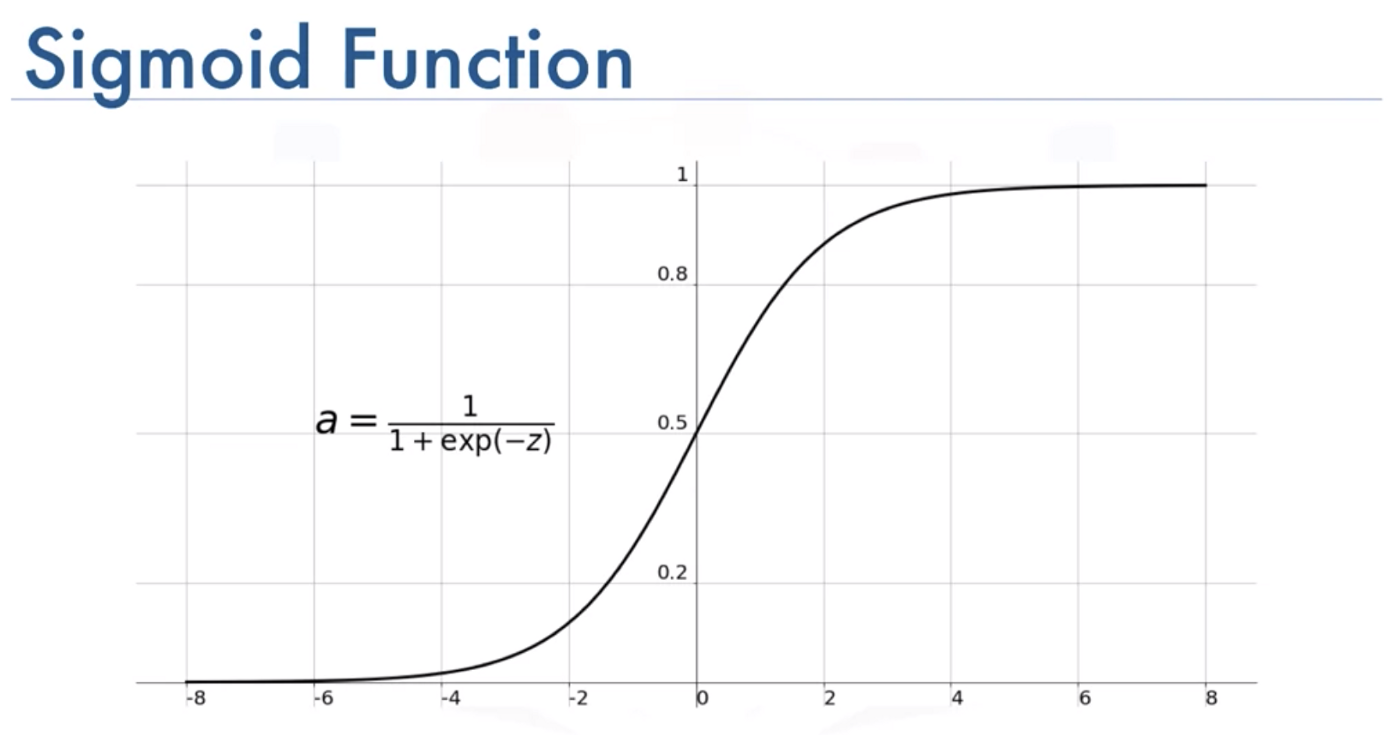
# What is Logistic Regression?

How do we make choices as human beings? We consider the options in front of us and the one that has a higher probability of being the correct option is chosen. Machine Learning models work similarly.

Let’s take a quick example for you to visualize this. Let’s say we want to predict if a high school student is going to be admitted to an elite college. The possible outcomes are that the student either will get admitted or will not be admitted. This sort of binary outcome of yes or no can be determined using a logistic regression model. To be more adept at understanding a logistic regression model, we need to understand the mathematics behind the logistic regression classifier. One quick point here, logistic regression is a supervised learning classification model that is used to predict a target variable's probability.

## Logistic Function (Sigmoid Function)

A good way to remember the sigmoid function is as an “S”-shaped curve. It looks something like this.

  
[Source](mailto:https://medium.com/@toprak.mhmt/activation-functions-for-deep-learning-13d8b9b20e)

### Why are we discussing this for logistic regression?

The sigmoid function is also called the squashing function. The aim of the function is to squash the set of all real numbers between 0 and 1. So the range of the function is (0,1). If we were to pass any number between -∞ and +∞ to the sigmoid function, the result would be bounded between only two numbers, i.e., 0 and 1.

Now we are getting closer to our understanding. Whatever the input is, the output can only be yes or no. Now, how do we decide whether the output should be 0 or 1? For this, we use a **threshold**.

  
[Source](mailto:https://www.javatpoint.com/logistic-regression-in-machine-learning)

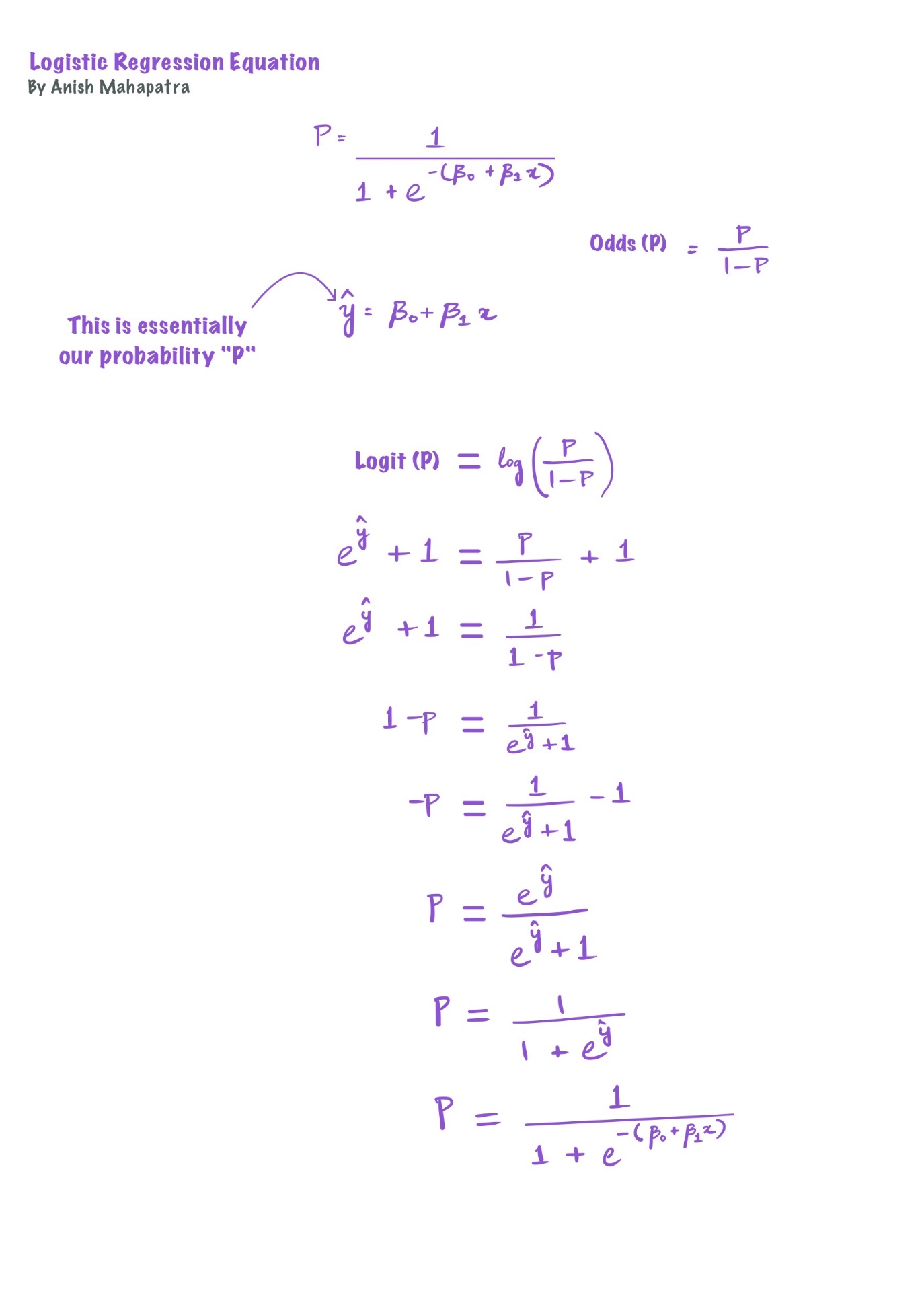
A threshold value indicates that all values that are equal to or above the threshold are mapped to one class and all the other values are mapped to the other class. When classification problems have an imbalance, tuning the threshold parameter can help to improve the performance of the model.

The default value of the threshold for the logistic regression classifier is 0.5, and tuning this particular hyperparameter is known as ***threshold moving***. Here is an interpretation that will help your understanding:

* If the prediction < 0.5, then it is Class 0
* If the Prediction is >= 0.5, then it is Class 1

## Logistic Regression Equation

Now that you have a solid understanding of the intuition behind the sigmoid function, let’s understand how the logistic regression equation is derived.



[Source](https://www.linkedin.com/in/anishmahapatra/)

# Three types of Logistic Regression

There are primarily three types of logistic regression models, let us go through them one by one.

## Binary Logistic Regression

A binary logistic regression is a type of logistic regression model where a response variable can belong only to two categories. This means that there are just two possible outcomes.

**Example** – Whether a bank is going to lend money to a customer or not. The outcome for this is binary, it’s either a yes or a no.

Another example can be assessing the risk of cancer for a patient, where the outcomes can be high or low.

## Multinomial Logistic Regression

Logistic regression is one of the most commonly used models and multinomial logistic regression is used where we can set multiple thresholds in the range of (0,1) to assess which class a variable can belong to. Let us understand a little more about logistic regression for multiclass classification with the help of an example.

As we discussed earlier, logistic regression is generally used for two-class classification. Multinomial logistic regression to perform multiclass classification (3 or more classes) –is a modified version of logistic regression that predicts more than one class for each input example.

**Example** – Predicting if a high school student will go to college, not go to college or into the workforce can be one use case.

In the case of image recognition, logistic regression can be used to determine if an image is that of a leopard, cheetah or lion.

## Ordinal Logistic Regression

Ordinal logistic regression is similar to multinomial logistic regression except for the fact that there is an ordering of classes that are required. This might mean that one class is ranked higher than another class.

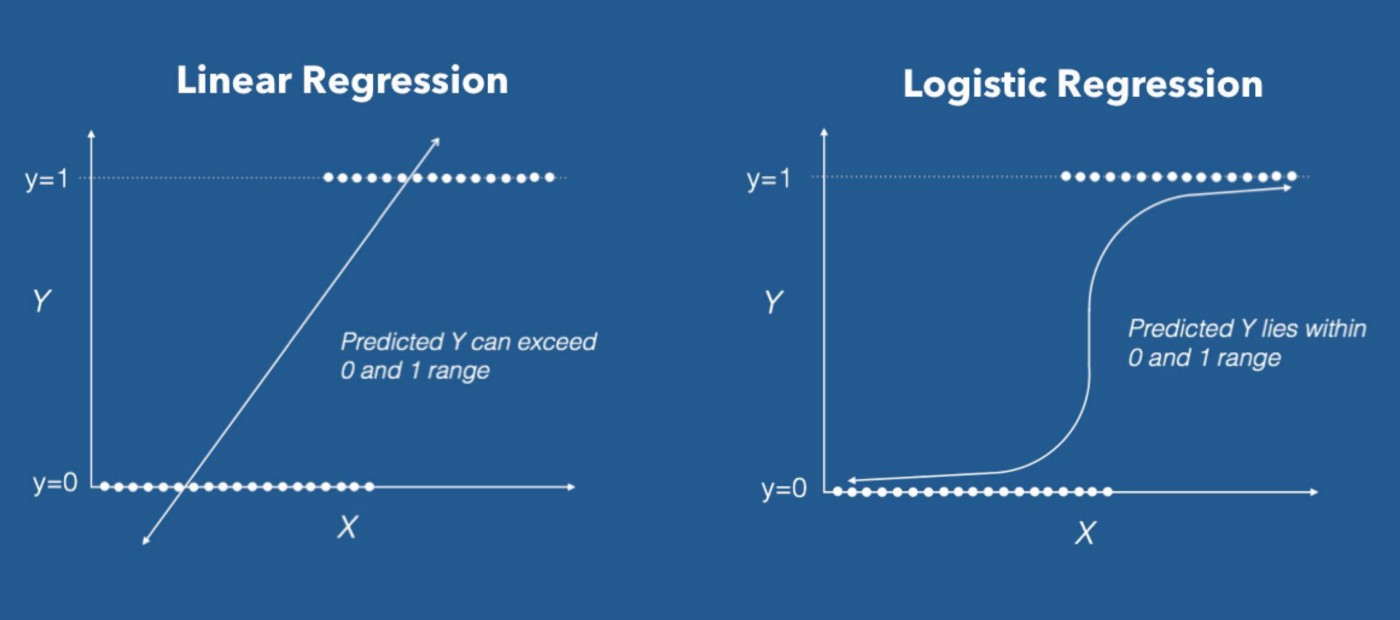
**Example** -

Based on multiple factors, we can decide the income range of a person to be in any of 5 categories:

* Category 1: ₹0 < Salary <= ₹5,00,000
* Category 2: ₹5,00,000 < Salary <= ₹10,00,000
* Category 3: ₹10,00,000 < Salary <= ₹20,00,000
* Category 4: ₹20,00,000 < Salary <= ₹30,00,000
* Category 5: ₹40,00,000 < Salary

# Logistic Regression Hypothesis\*\*

The Logistic regression hypothesis is that the cost function tends to limit the output between 0 and 1. Linear functions fail to represent this as they can go below 0 or, above 1. According to the hypothesis of logistic regression, this is not possible.



[Source](https://www.datacamp.com/tutorial/logistic-regression-R)

# How does Logistic Regression Algorithm work?

Logistic regression in ML analyzes the latent relations between variables and using the sigmoid function, assigns probabilities to discrete outcomes using the Sigmoid function. The Sigmoid function converts any real number to a space between 0 and 1.

It is a technique that helps to spell out the probability of an event that is going to happen in the future. The model can take in historical data, learn from the data and predict the likelihood of the event happening. The main principle of logistic regression is to use a model to predict the probability that an outcome will occur.

## Sigmoid function

The sigmoid function helps squeeze the output in the range of 0 and 1. The reason why we use the sigmoid function is that, based on the threshold, we can snap the output to either 0 or 1.



[Source](https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/)

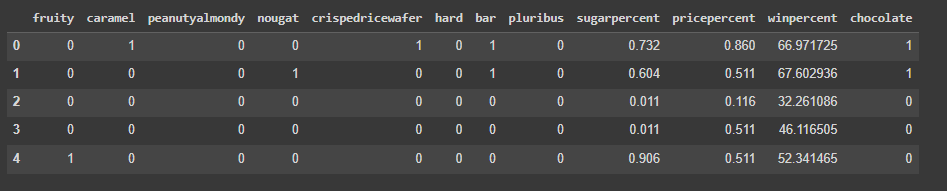
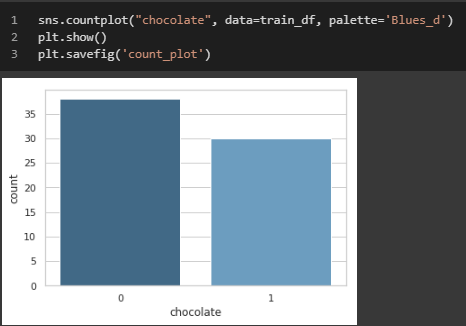
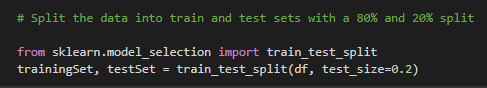
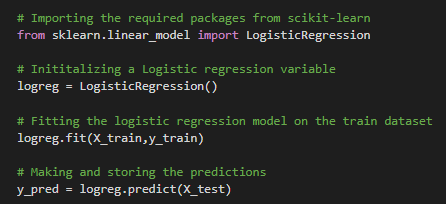
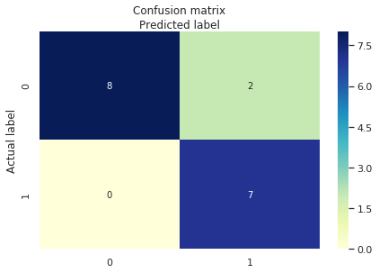
# When and Where to use Logistic Regression?

The logistic regression algorithm in machine learning is used to solve problems of classification. In simple terms, if there are discrete outputs, preferably binary outputs (yes or no), then logistic regression can be used. The logistic algorithm is used in almost every domain:

* Healthcare: Logistic regression can be used to aid doctors in understanding if a patient has a particular illness or not. A popular example is that logistic regression can gauge if a tumour is likely benign or malignant.
* Marketing & Sales: In sales roles, logistic regression can be used to predict if a lead is likely to convert or not. In marketing roles, a logistic regression algorithm can be used to gauge if a particular target audience is likely to click on the link or not.
* Finance: In banking or the finance industry, logistic regression can be used to know if a transaction is likely fraudulent or not.

# Building Logistic Regression Model

There are many ways that one can build a logistic regression model. Following are some of the steps over an example that can help you understand the flow of how to build a logistic regression model. Please feel free to comment below in case you would like the code as well!

* Decide the dataset that you are going to use and proceed to load the dataset onto your IDE of choice. You can find the dataset that we are using [here](https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking) in this example on GitHub   
  
* Explore the dataset thoroughly using Exploratory data analysis techniques such as univariate analysis, bivariate analysis, outlier analysis, and distribution analysis. In this case, our target variable here is “chocolate”. The aim of the data is to decide based on multiple factors if the given candy is considered chocolate or not.
* Make sure the dataset is compliant with the assumptions of logistic regression (we will discuss more on the assumptions of logistic regression in a later section)
* Visualize the dataset as much as possible, this will give you a lot of insights. Here, we have presented a simple visualization for your reference.   
     
  
* Split the data into a training and testing dataset   
  
* Fit the logistic regression model on the training dataset   
  
* Evaluate the model using evaluation metrics of a classification problem. We will do this using a confusion matrix.   
  

# Applications of Logistic Regression

Let us look at a few real-life examples of how logistic regression can be applied in various industries.

* **Medicine**: Based on multiple data points collected from the user, it is possible to predict whether a patient is likely to have a heart attack or not. A logistic regression model can help us do this.
* **Email**: A logistic regression model can help us identify if an email that is sent is spam or not. This can be done by analyzing the text in the email, the sender, the country of origin and multiple other factors.
* **Credit Card Fraud**: It is mission critical for multiple companies to identify whether a transaction is fraudulent or not. A transaction can be classified as fraudulent or not fraudulent based on this analysis.
* **Hotel Booking**: A giant hotel booking site, Booking.com, revealed in ti’s HighLoad++ Siberia conference that it uses logistic regression to predict whether a user will change their selected journey date or not. Based on this, the company can perform A/B Testing to determine if it should change the interface for one class of users.
* **Gaming**: Gaming Companies like Tencent collect millions of data points from their users. Based on this information collected, Tencent recommends its users the gaming equipment that they are likely to buy.

# Pros and Cons of Logistic Regression

In this section, we will discuss the pros and cons of logistic regression.

## Pros of Logistic Regression

* **Highly interpretable & less computation**: Logistic regression is a widely used machine learning algorithm that does not require much computation. It is also highly interpretable.
* **Easy to implement**: It is one of the most effective algorithms to train and implement. It performs well when the data is *linearly-separable*.
* **Scalable**: It is easy to scale the input features of logistic regression, it does not require additional tuning.
* **Gives importance of features**: Logistic regression gives a direction of the association, whether it’s positive or negative. It also tells us how important a predictor is based on the coefficient size. Basically gives feature importance.
* **Less inclined to overfit**: Unless the data has very high dimensions, logistic regression is less inclined to overfit.

## Cons of Logistic Regression

* **Bound by linearity**: Logistic regression works well with data that is linearly separable as it forms linear boundaries. It assumes that the relationship between the dependent and independent variables is linear.
* **Cannot obtain complex relationships**: There are more powerful algorithms like neural Networks that can perform better than logistic regression as the more advanced algorithms can find complex and latent relationships between the features.
* **Does not allow multicollinearity**: No multicollinearity r average multicollinearity is allowed between the features of the input dataset.
* **Discrete output set**: Logistic regression only allows for a discrete output. This is a drawback.

# Making predictions with Logistic Regression

Logistic regression is a supervised machine learning classification model. It is used to primarily predict the class based on multiple predictor variables from the historical data present. Logistic regression is of three types – binary, multinomial and ordinal. It belongs to a family called Generalized Linear Model. With logistic regression, gradient descent can be used as an iterative optimization algorithm to find the minima of a differentiable function.



[Source](https://www.baeldung.com/cs/gradient-descent-logistic-regression)

Let us now look at some of the assumptions of Logistic regression.

## Logistic Regression Assumptions

* **Binary discrete output set**: Logistic regression is made for binary output classification. When modifications are made, it is possible to perform multi-class or ordinal classification.
* **Sensitive to outliers**: It assumes that there is no error in the output set. So, it would make sense to remove outliers and misclassified points to make sure we get the best use out of the logistic regression algorithm.
* **Follows a Gaussian Distribution**: Logistic regressions assume a linear relationship between the dependent and independent variables. Data transformations such as box-cox transformation, log or root may be needed to get this relationship.
* **Can Fail to converge**: If the dataset has highly correlated variables or the data is very sparse, it is possible that using a maximum likelihood estimator, the logistic regression algorithm might fail to converge.
* **Sensitive to Multicollinearity**: The model can overfit if the input features are highly correlated.

# Logistic Regression Best Practices for 2022

Let’s discuss the best practices for logistic regression.

* Focus on dependent variables to ensure that the model is consistent
* Ensure that the assumptions of logistic regression are met to avoid overfitting or underfitting
* It is important to evaluate the goodness of fit while evaluating the model
* Machine Learning interpretability can be derived based on the coefficients, also indicating the direction of relation with the output variable
* Validation of observed results with a subsample is critical as this can reduce the bias in the model and help validate the model

# Logistic vs Linear Regression

A fun fact here, even though it is called Logistic “Regression”, it is used for classification and not regression. It is called ‘Logistic Regression’ because the underlying principle is similar to that of Linear Regression. The ‘Logistic’ part of the name is derived from the ‘Logit’ function.

Let us see a few differences between linear regression and logistic regression

* Logistic regression is used to perform classification, whereas linear regression is used to perform regression
* The output of logistic regression is discrete, whereas the output of linear regression is continuous
* The aim of logistic regression is to fit the values in the S-shaped sigmoid curve, whereas linear regression is to find the best-fit line
* The loss function of logistic regression is the maximum likelihood estimator, whereas the loss function of linear regression is the mean squared error

# Conclusion

In this article, we have understood the ins and outs of logistic regression. It is a detailed article, so please feel free to go through the article multiple times and refer to the sources mentioned. In interviews as well, logistic regression is one of the most asked questions, so make sure to familiarize yourself with it. Even in real-world machine learning problems, the baseline model that is used for a lot of classification models is logistic regression. It is an extremely important topic that you have successfully completed. Remember, the best way to learn Data Science is to simply do it!

# FAQs

### Q. Why is logistic regression used in Machine Learning?

Logistic regression is a supervised learning classification model. It is used to predict in a binary (two classes) fashion the likelihood of one class or the other. It is used as a baseline classification model in multiple use cases and has applications throughout the industry. It is a simple and efficient method that also scales well and is quite interpretable.

### Q. Is logistic regression a model or algorithm?

In Data Science, the words “model” and “algorithm” tend to be used interchangeably. Both of them in most situations denote the same thing. So, a logistic regression model is the same thing as a logistic regression model.

### Q. Why do we use logistic regression instead of linear regression?

Linear regression is used to perform regression and the output is a continuous variable. When we squeeze this where the output set lies between 0 and 1 and is used to perform binary classification, it is known as logistic regression.

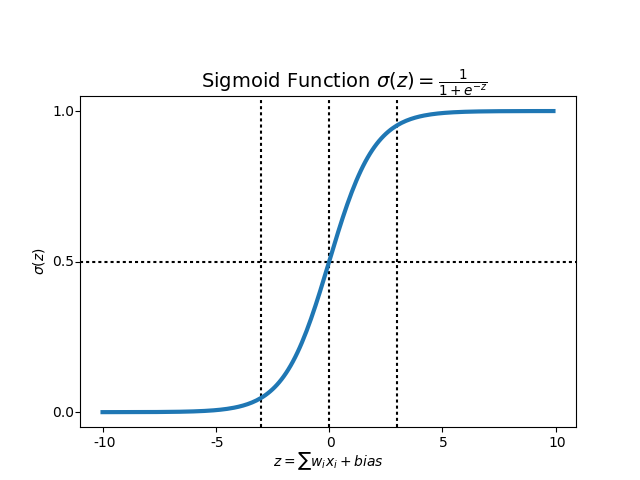
We use logistic regression instead of linear regression when the task at hand class for a discrete output and the prediction needs to be a classification (Yes/No).

### Q. Why logistic regression is the best model?

Logistic regression is a great baseline model. This is because it is easy to interpret, scales well and can be implemented quickly and effectively. The fundamental understanding of logistic regression is widely used across the industry.

### Q. What algorithm is used in logistic regression?

The logit function is used in logistic regression. It is a special case of a generalized linear model.



# Author

**Anish Mahapatra**   
*Senior Data Scientist*   
LinkedIn: [https://www.linkedin.com/in/anishmahapatra](about:blank) ([LinkedIn](about:blank))   
Email: anishmahapatra01@gmail.com