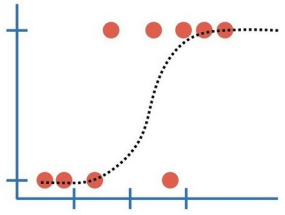
Logistic Regression

Tushar B. Kute,
http://tusharkute.com



Logistic Regression

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
- Like all regression analyses, the logistic regression is a predictive analysis.
- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
- Remember: though the name of algorithm carries regression, it is used for classification.

Type of Logistic Regression

- Binary Logistic Regression
  - The categorical response has only two 2 possible outcomes. Example: Spam or Not.
- Multinomial Logistic Regression
  - Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan).
- Ordinal Logistic Regression
  - Three or more categories with ordering. Example: Movie rating from 1 to 5.

Logistic Regression

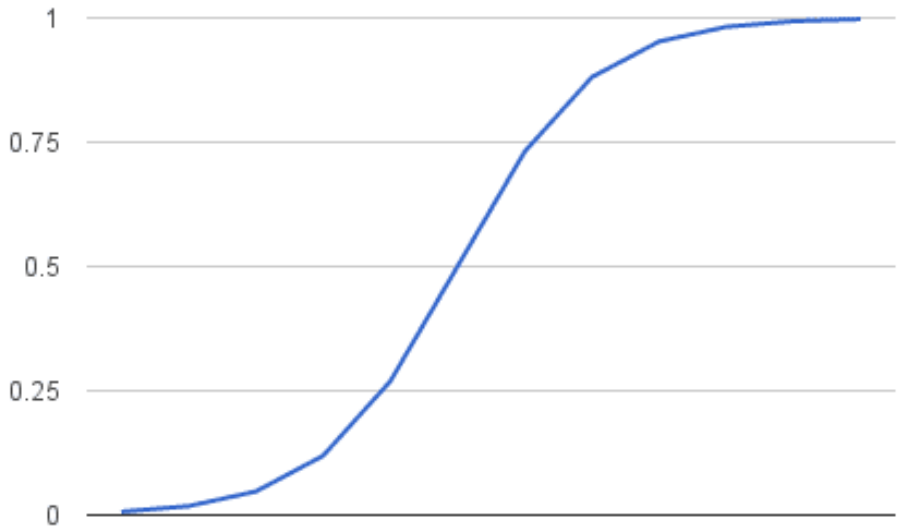
- Logistic regression is named for the function used at the core of the method, the logistic function.
- The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment.
- It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Ref.: https://machinelearningmastery.com

Logistic Regression

- e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.



Logistic Regression

- Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).
- A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.
- Below is an example logistic regression equation:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Logistic Regression – Predictions

- Logistic regression models the probability of the default class (e.g. the first class).
- For example, if we are modeling people's gender as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person's height, or more formally:

$$P(\text{gender=male}|\text{height})$$

- Written another way, we are modeling the probability that an input (X) belongs to the default class (Y=1), we can write this formally as:

$$P(X) = P(Y=1|X)$$

Logistic Regression – Predictions

- Let's say we have a model that can predict whether a person is male or female based on their height (completely fictitious). Given a height of 150cm is the person male or female.
- We have learned the coefficients of b0 = -100 and b1 = 0.6. Using the equation above we can calculate the probability of male given a height of 150cm or more formally P(male| height=150). We will use EXP() for e, because that is what you can use if you type this example into your spreadsheet:

$$y = e^{(b0 + b1*X)} / (1 + e^{(b0 + b1*X)})$$

$$y = \exp(-100 + 0.6*150) / (1 + EXP(-100 + 0.6*X))$$

$$y = 0.000453978687$$

Or a probability of near zero that the person is a male.

Logistic Regression – Predictions

- In practice we can use the probabilities directly. Because this is classification and we want a crisp answer, we can snap the probabilities to a binary class value, for example:

  0 if p(male) < 0.5

  1 if p(male) >= 0.5

Preparing data

- Binary Output Variable: This might be obvious as we have already mentioned it, but logistic regression is intended for binary (two-class) classification problems. It will predict the probability of an instance belonging to the default class, which can be snapped into a 0 or 1 classification.
- Remove Noise: Logistic regression assumes no error in the output variable (y), consider removing outliers and possibly misclassified instances from your training data.

Preparing data

- Gaussian Distribution: Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output.
- Remove Correlated Inputs: Like linear regression, the model can overfit if you have multiple highly-correlated inputs.
- Fail to Converge: It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge. This can happen if there are many highly correlated inputs in your data or the data is very sparse (e.g. lots of zeros in your input data).

What we know?

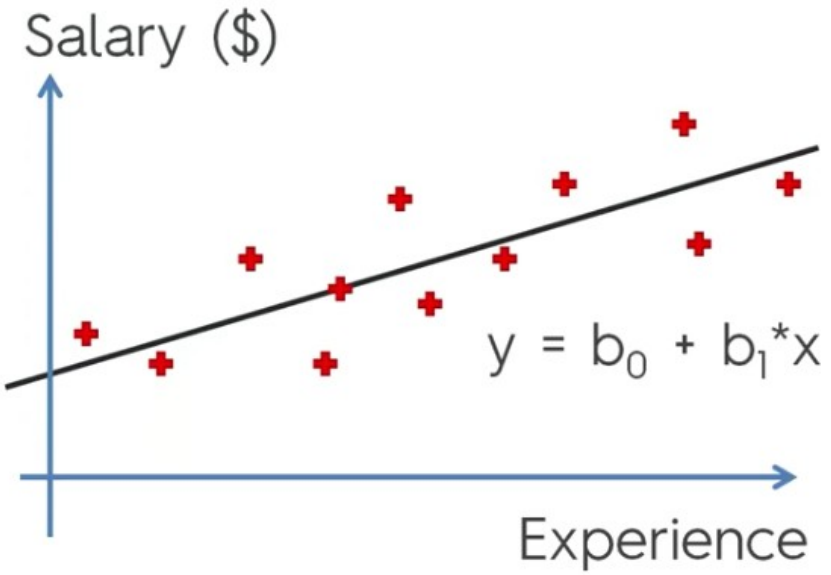
Linear Regression:

- **Simple:**

$$y = b_0 + b_1 * x$$

- **Multiple:**

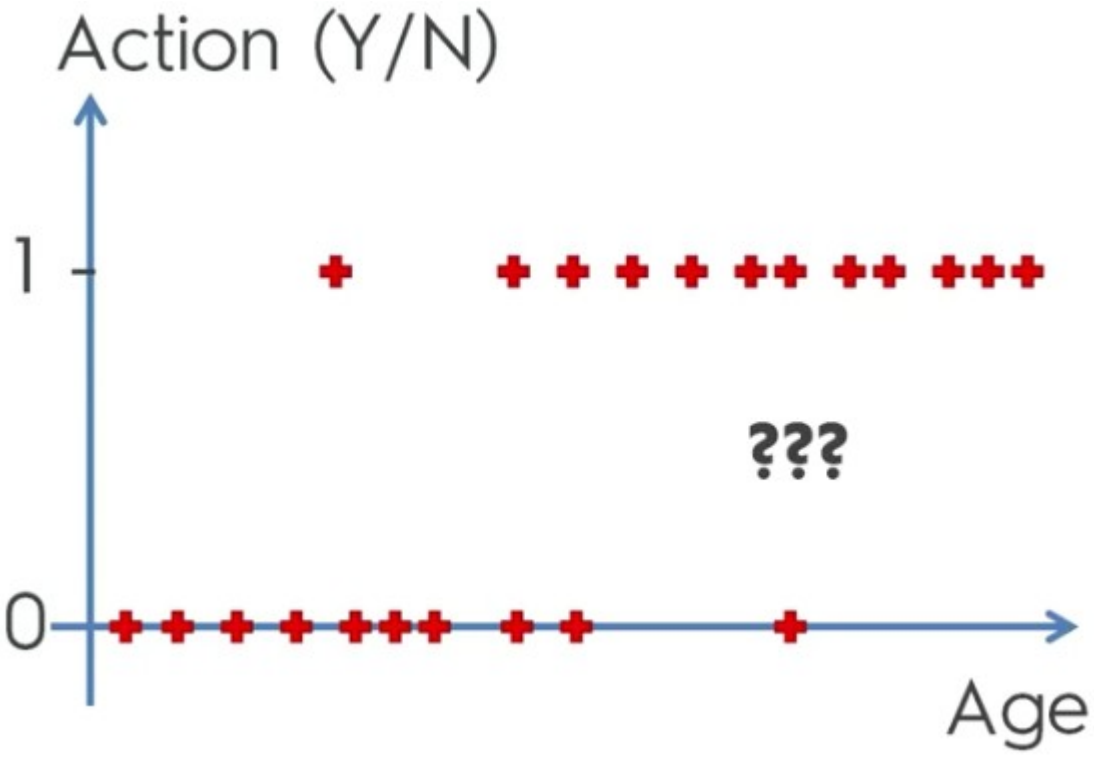
$$y = b_0 + b_1 * x_1 + \ldots + b_n * x_n$$



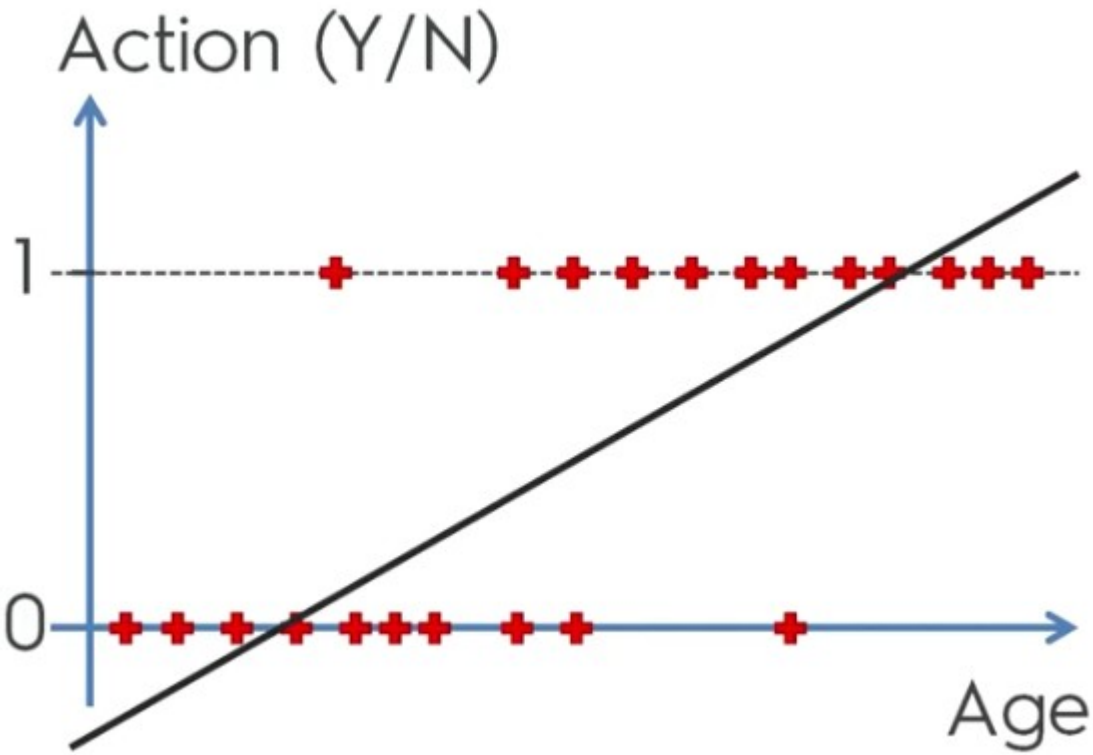
Example Reference: Super Data Science

A new problem

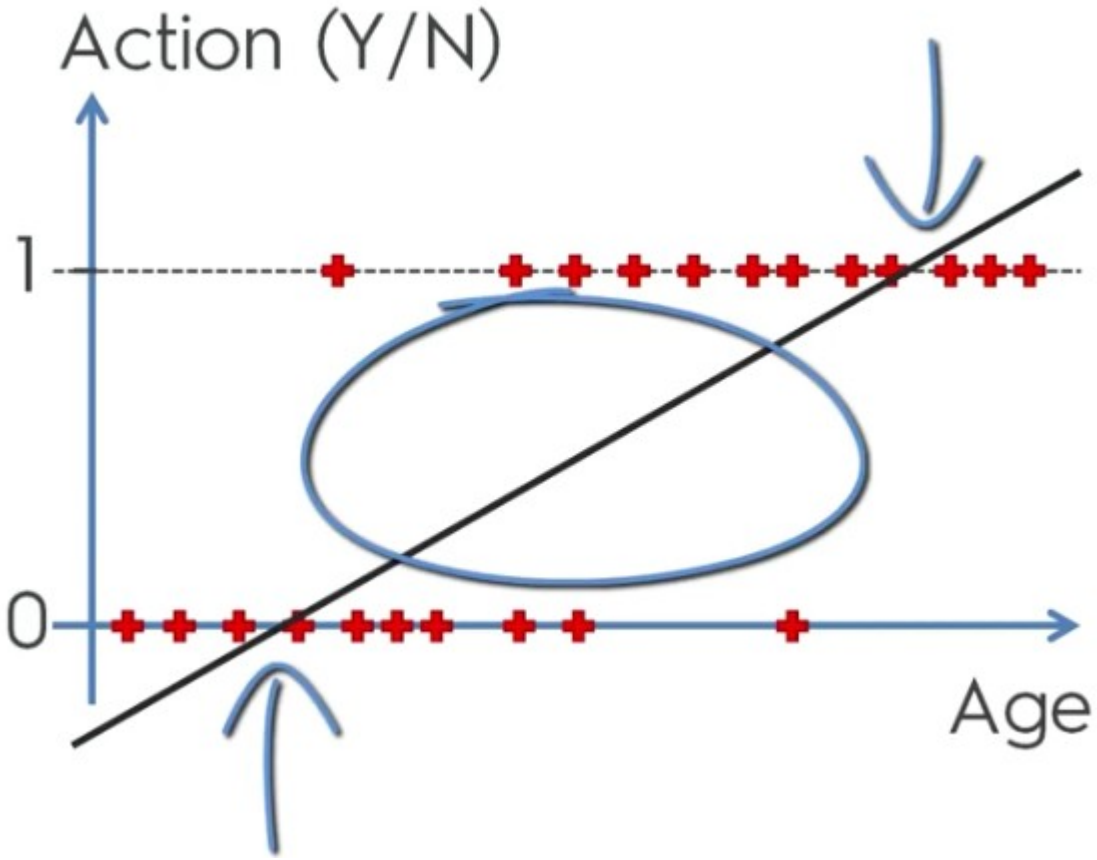
A company has provided an offer by email to their customers.



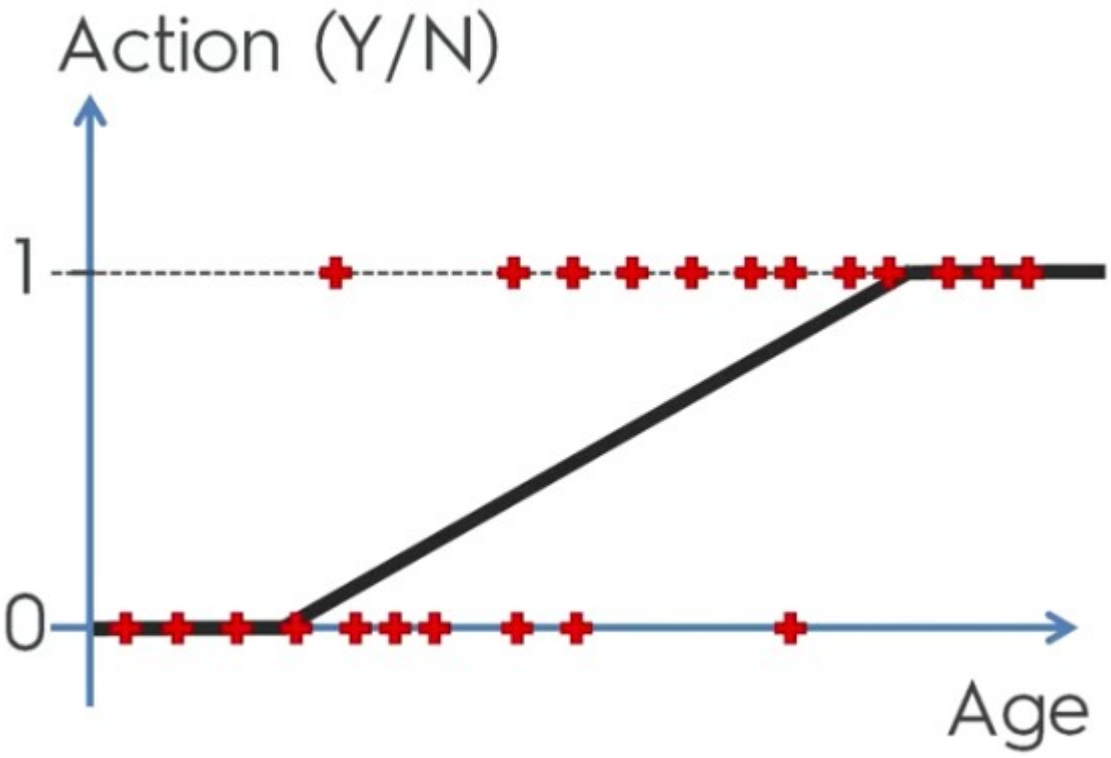
Apply Linear Regression



Apply Linear Regression



Apply Linear Regression



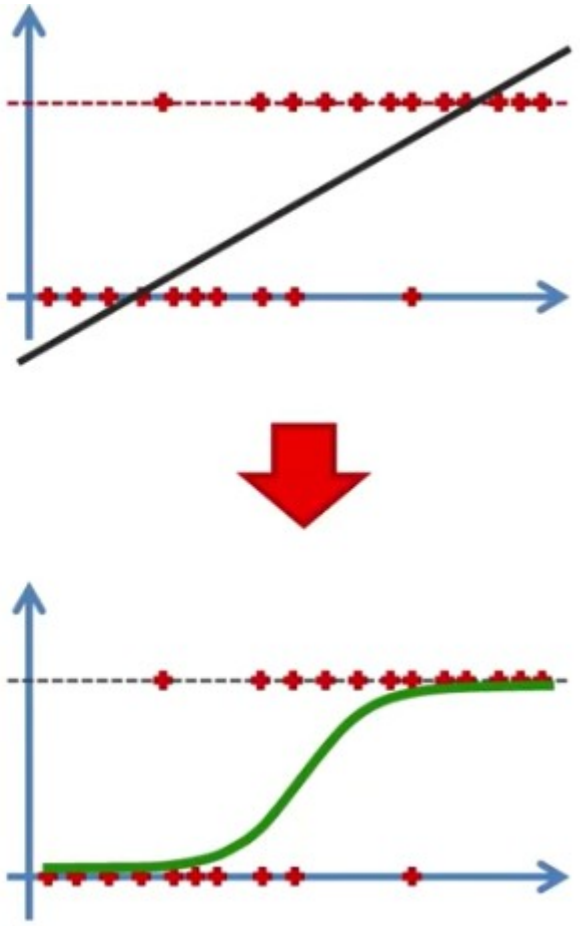
Logistic Regression

$$y = b_0 + b_1 * x$$

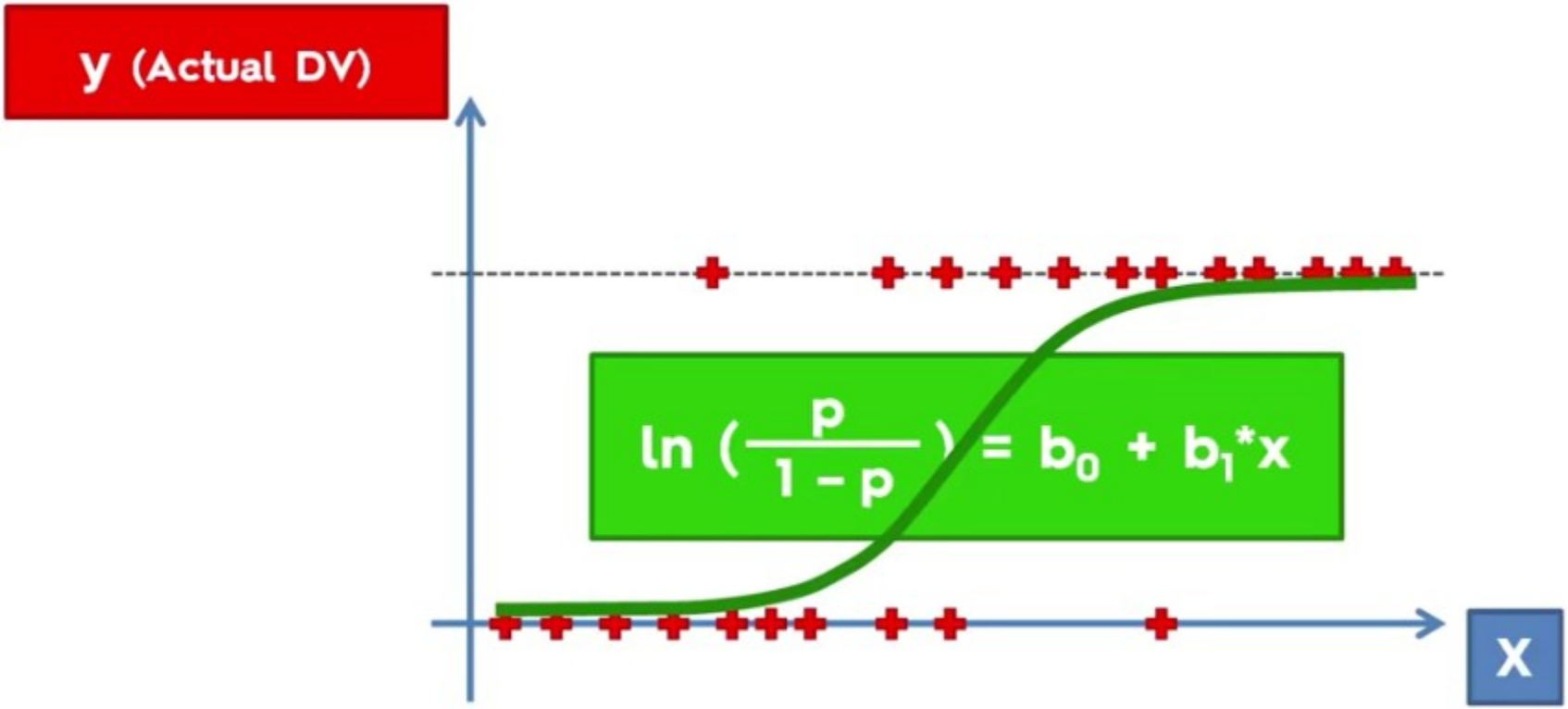
Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

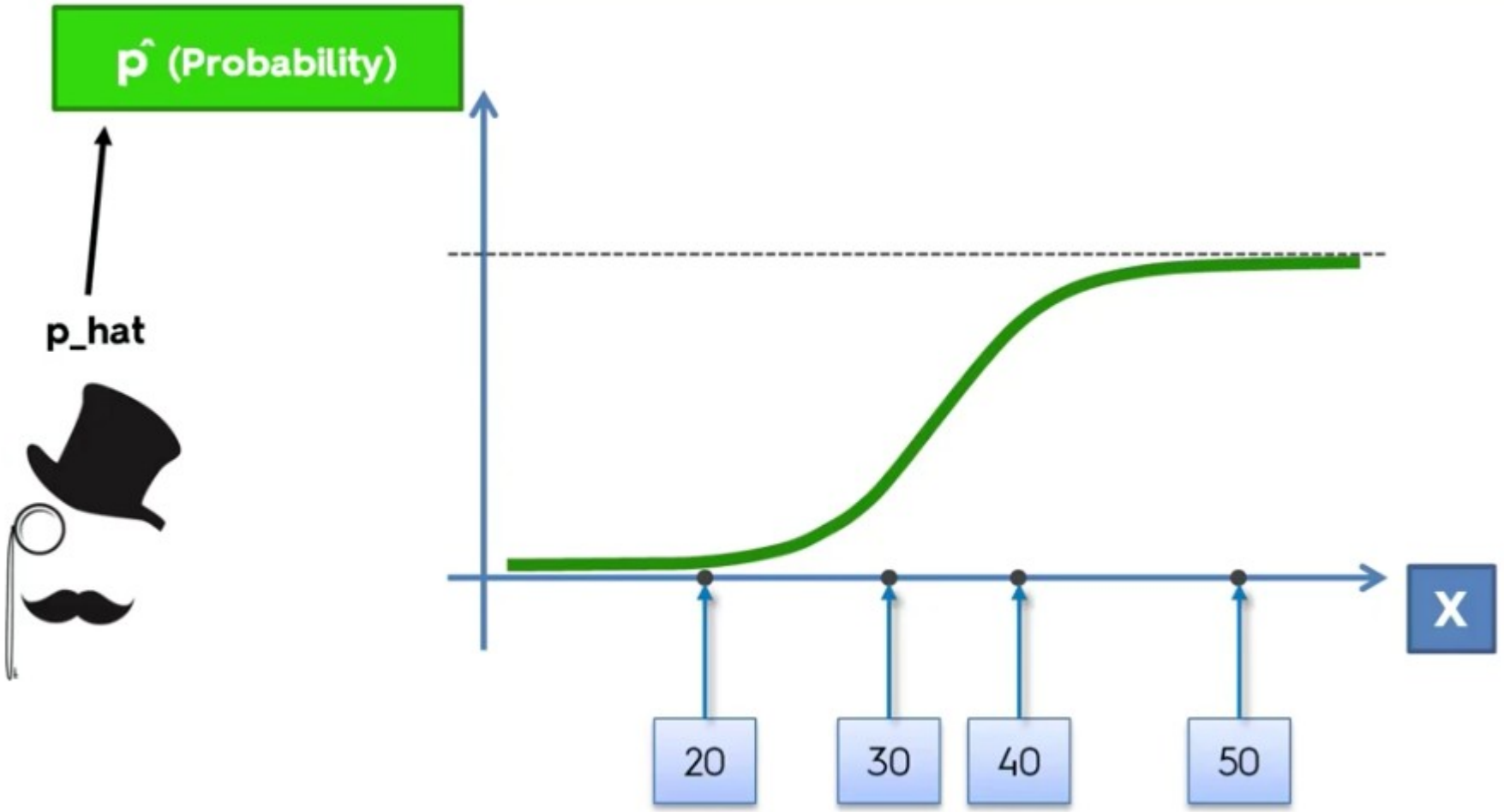
$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$



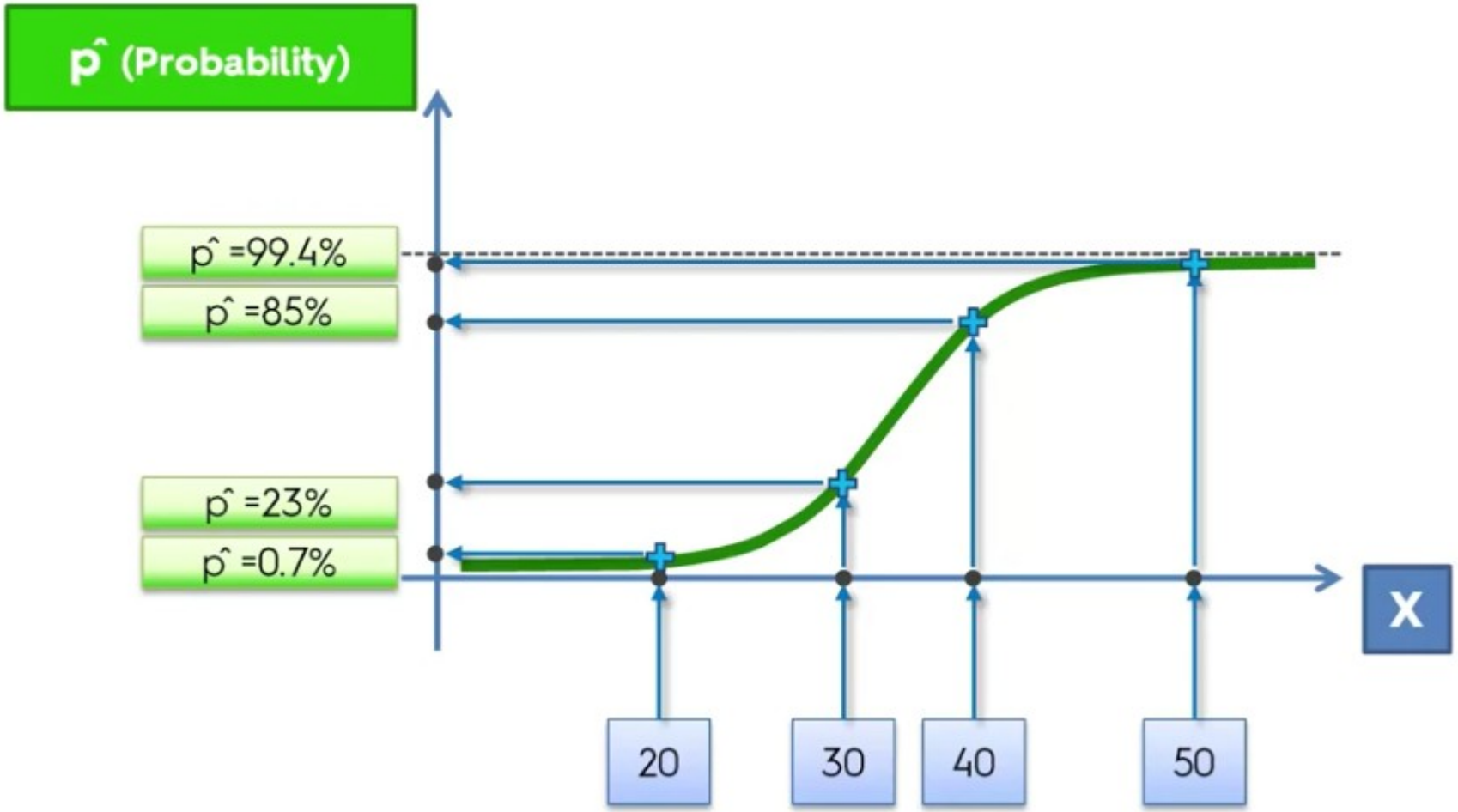
Logistic Regression – Logit Function



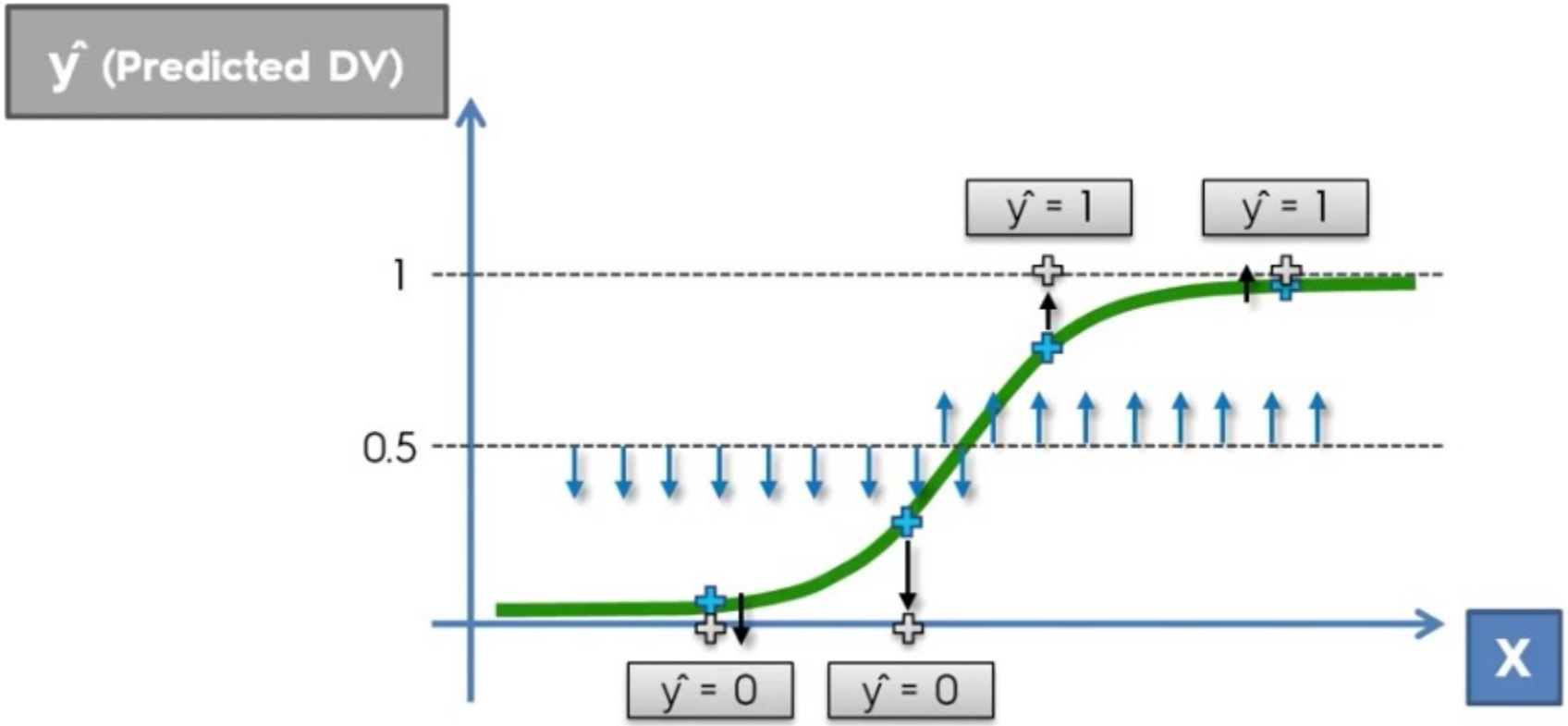
Logistic Regression



Logistic Regression – Probabilities



Logistic Regression – Prediction



Advantages

- Logistic Regression performs well when the dataset is linearly separable.
- Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.
- Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).
- Logistic regression is easier to implement, interpret and very efficient to train.

Disadvantages

- Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.
- If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit.
- Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set.

Useful resources

- www.superdatascience.com
- www.mitu.co.in
- www.pythonprogramminglanguage.com
- www.scikit-learn.org
- www.towardsdatascience.com
- www.medium.com
- www.analyticsvidhya.com
- www.kaggle.com
- www.stephacking.com
- www.github.com

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com