DATA SCIENCE
DREAM JOB

Logistic Regression

Table of Contents

1a. Overview

Logistic Regression is derived from the Logit Transformation

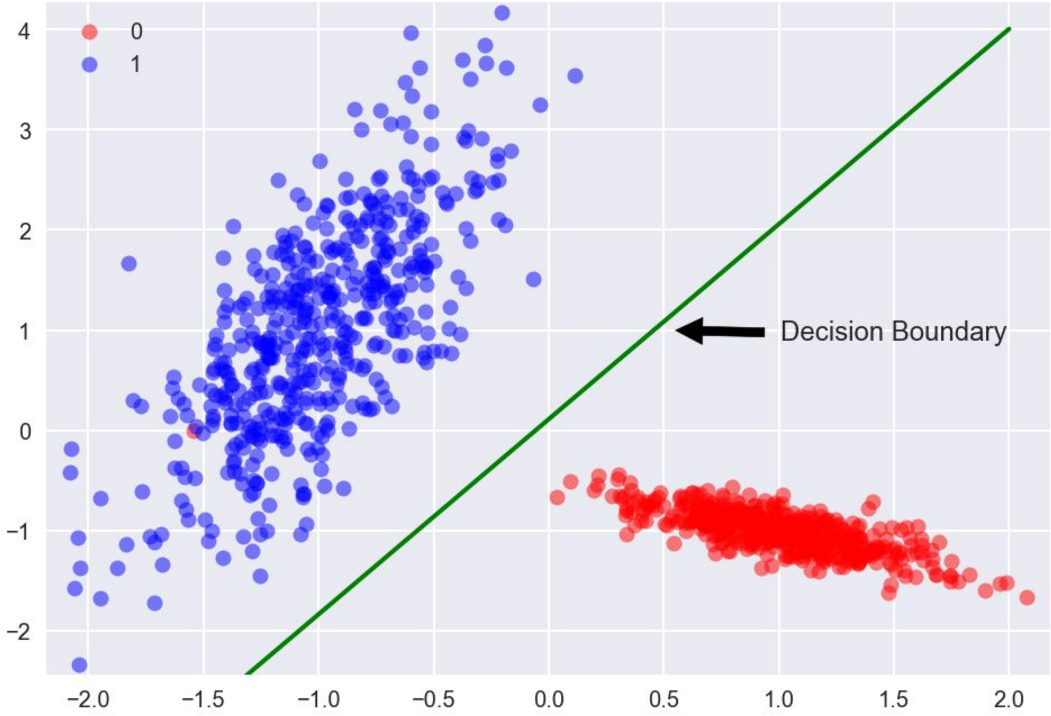
And is an algorithm that learns a model for binary classification

Logistic Regression Concepts:

- Predicts binary outcomes
- Predicts probability scores
- DV must have discrete values (i.e. True/False)
- Uses a Logit Transformation on the DV to fit linear model

Example

Fit a line that **best separates** the classes



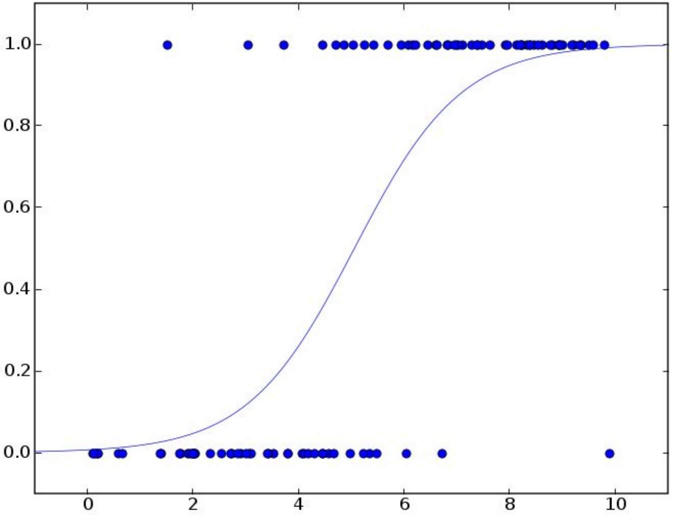
Source:
https://camo.githubusercontent.com/f663cd4f29335972950dded4d422c07aeee8af55/687474707
33a2f2f63646e2d696d616765732d312e6d656469756d2e636f6d2f6d61782f313630302f312a344
7306773753932725068 4e2d636f3970763150354140327 82e706e67

Example (2)

Logistic Regression allows you to make:

i. Soft Predictions (% Probability)

ii. Hard Predictions (0’s, 1’s)



Predicting Credit Default:

Name	Predicted Default Probability	
Bob	0.85	
Henry	0.76	
Mary	0.53	
Paul	0.32	**Threshold**: 0.15
Maria	0.14	
Sonny	0.12	
Ryan	0.07	
Howard	0.04	

Custom Threshold

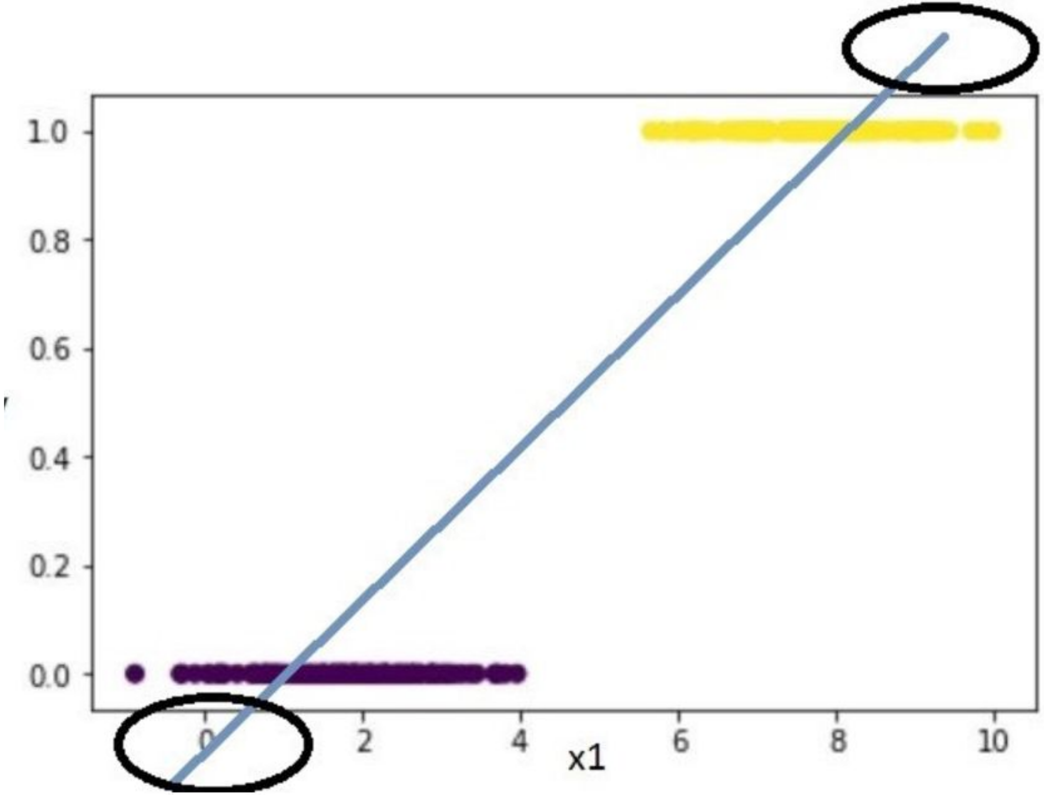
1b. Logistic Regression Model

- Uses **Logit Transformation** on DV to fit a regression model
- DV has two (binary) outcomes

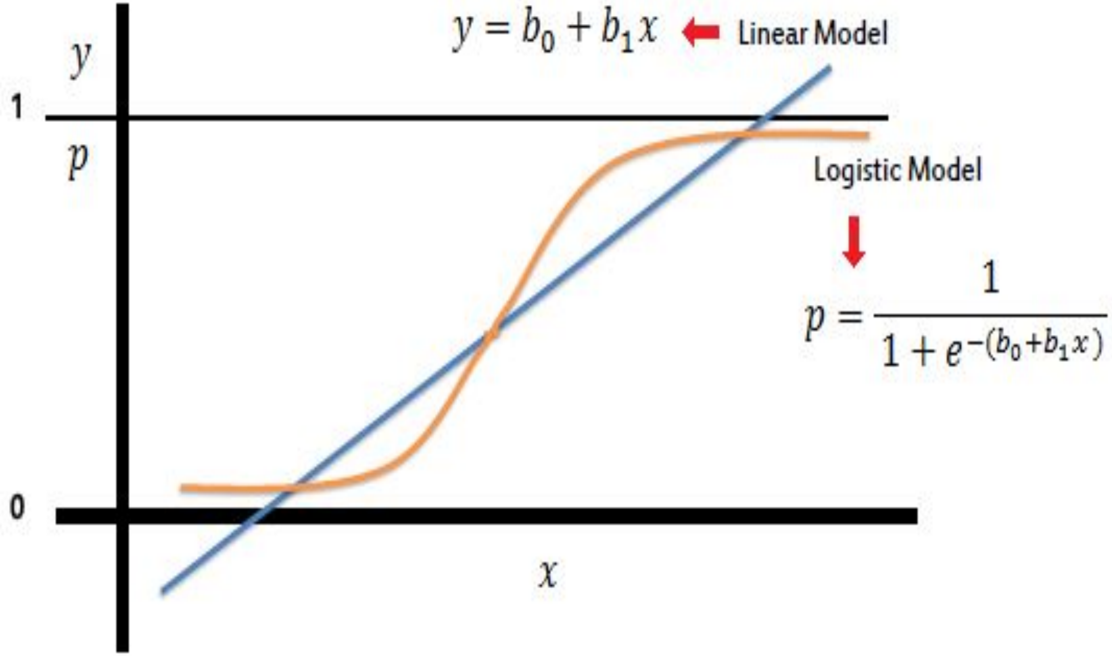
$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Logit Transformation

The Problem with Linear Regression...



Logistic Regression vs Linear Regression



1c. Important Concepts & Terminologies

Logistic Regression models the **probability** of an event - rather than a measure.

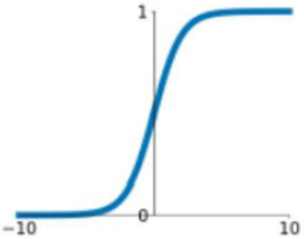
1. Probability
2. Odds
3. Logit/Log Odds
4. Sigmoid
5. Log Loss

- Probability = $\dfrac{\text{\# events}}{\text{\# subjects}}$
- Odds = $\dfrac{\dfrac{\text{\# events}}{\text{\# subjects}}}{\dfrac{\text{\# non-events}}{\text{\# subjects}}} = \dfrac{\text{probability}}{(1 - \text{probability})}$
- **Odds = p / (1 - p)**

  [ratio of two probabilities]

Sigmoid

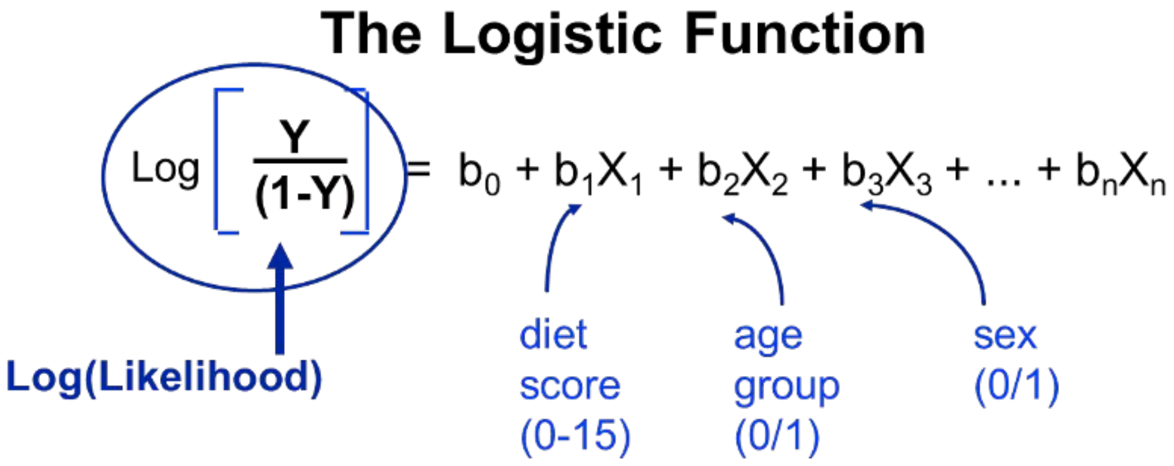
$$\sigma(x) = \frac{1}{1+e^{-x}}$$



$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right],$$

2. Understanding the Algorithm

- Logistic Regression models the **PROBABILITY** of an event
- Probabilities range from (0’s to 1’s)
- Requires a **LOGIT** transformation on DV (hence the name Logistic Regression)
- Still considered a Linear Model because of Input Parameters



2. Understanding the Algorithm

Linear Regression Equation:

Y = B0 + B1X1 + B2X2 +... + BkXk

Logit Regression Equation:

Log(Odds) = Ln(Y/1-Y) = B0 + B1X1 + B2X2 +... + BkXk

Why not just measure the probability?

EX) P = B0 + B1X1 + B2X2 +... + BkXk ????

How does it predict probability scores?

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\Downarrow$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$\Downarrow$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

2b. Cost Function - Log Loss

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Objective: Minimize the Log Loss Error

Logistic Regression uses the Log Loss Function (as opposed to OLS) to find optimal parameters

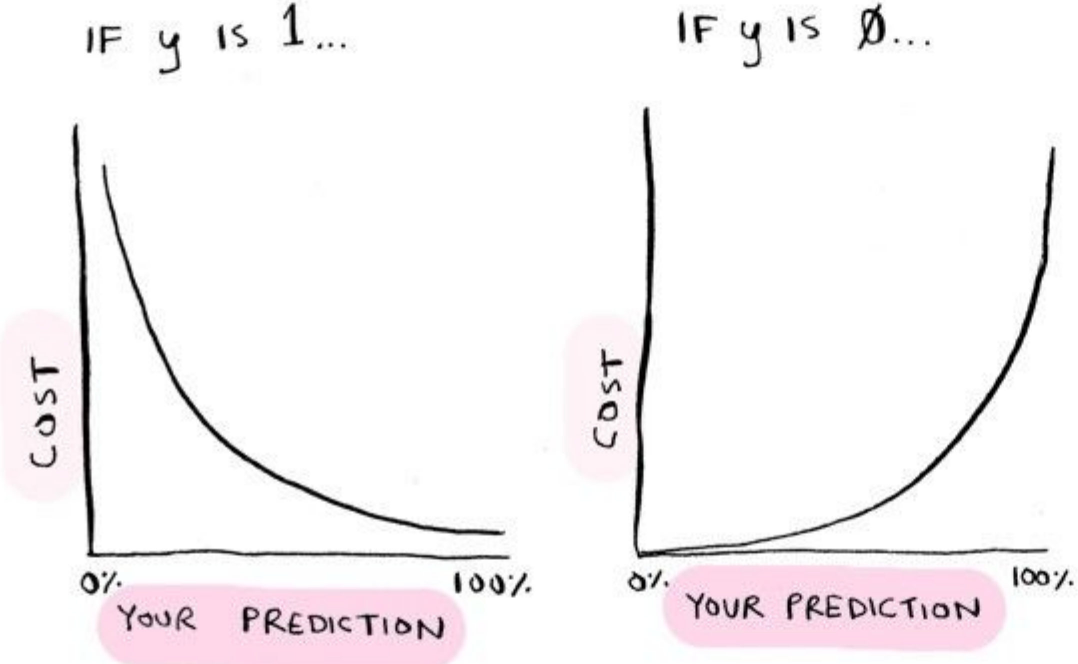
- **y(i)**: the DV (0 or 1)
- **h0(xi)**: predicted probability

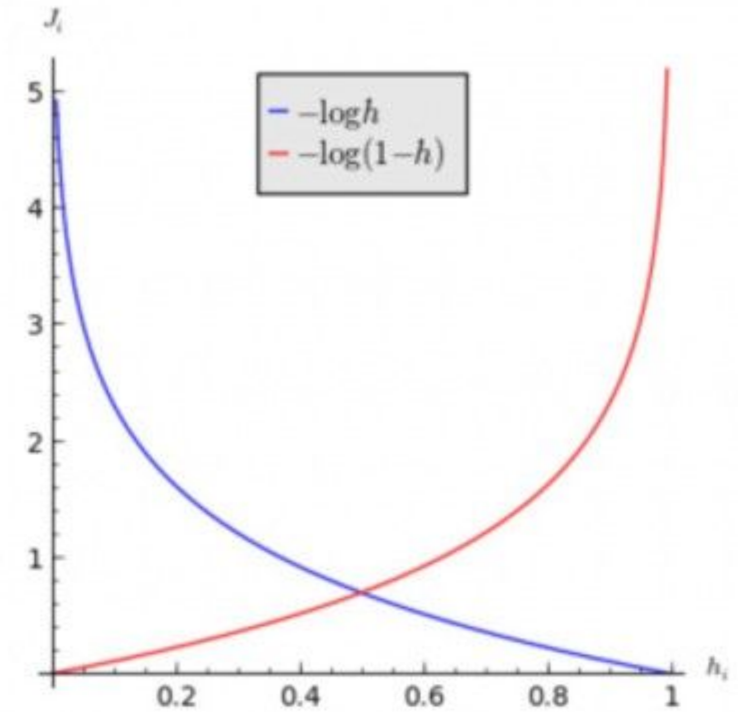
When **y=0** and probability is **LOW** → then low error

When **y=1** and probability is **HIGH** → then low error

Log Loss

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$





Source: https://cdn-images-1.medium.com/max/1200/1*DMnFxWJiATUgd3e32KNWOg.jpeg

3. Strengths and Limitations

💪 **Strengths:**

- Highly interpretable
- Fast training & predictions
- No model tuning required (except for regularization)
- No need to scale features
- Outputs probability scores
- Good baseline model

Limitations:

- Assumes a linear relationship
- Performance is not as great compared to other models
- Sensitive to outliers
- Can’t automatically learn feature interactions like tree-based methods

Summary

Logistic Regression is used for **binary classification** problem:

- It’s fast, simple, powerful, and is a great baseline model
- Uses a Logit Transformation on the DV to fit linear model
- Can make probability predictions
- Finds optimal parameters by minimizing log loss error

$$Ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$