EXPLORE AI
ACADEMY

Binary Classification & Logistic Regression

2304PTDS

December 2023

1 / **Objectives**

2 / **Binary Classification**

3 / **Logistic Regression**

4 / **To the code!**

Objectives

By the end of this session you should be able to:

- Understand binary classification
- Understand the basics of Logistic Regression as a classifier
- Distinguish between linear regression and logistic regression

Please do not copy without permission. © ExploreAI 2023.

1 / **Objectives**

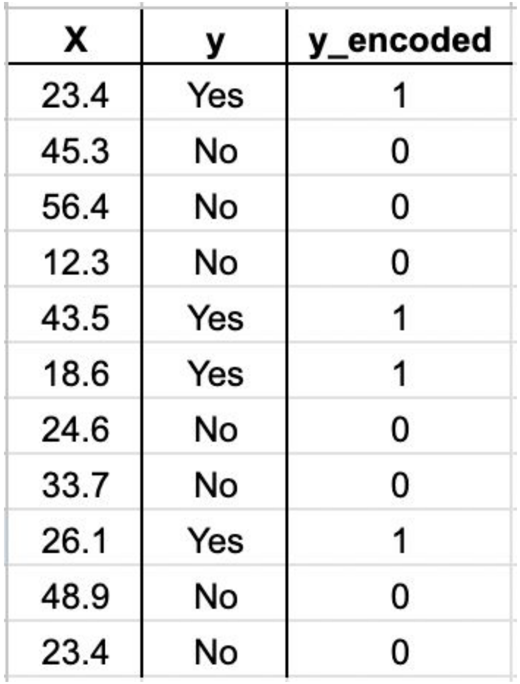
2 / **Binary Classification**

3 / **Logistic Regression**

4 / **To the code!**

What is Binary Classification?

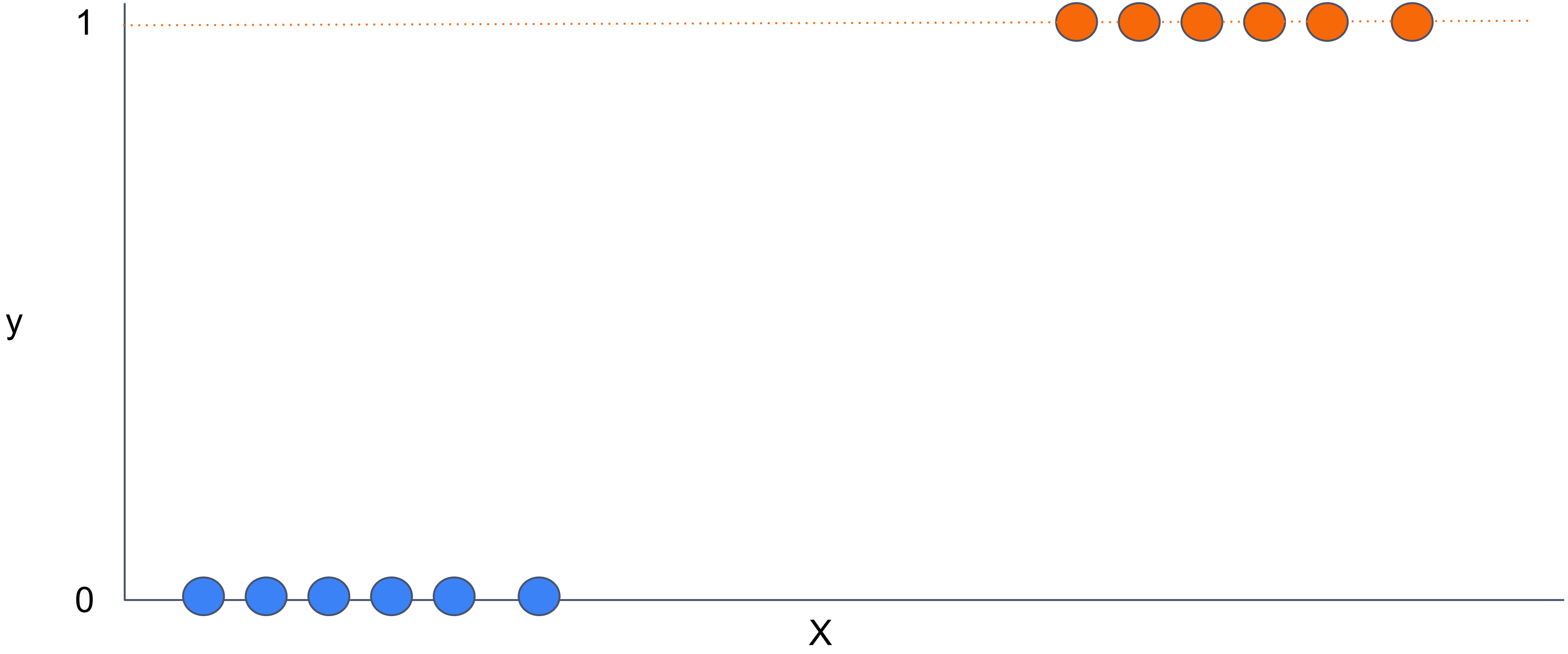
- Classification = when we are working with and trying to predict for categorical target
- Very simply put: classifying data into one of two categories
  - Yes/No,
  - Fraud/Not Fraud,
  - Spam/Not Spam,
  - True/False...
- Usually converted to 1s and 0s for use with machine learning models
- **But what happens when we plot them?**



X	y	y_encoded
23.4	Yes	1
45.3	No	0
56.4	No	0
12.3	No	0
43.5	Yes	1
18.6	Yes	1
24.6	No	0
33.7	No	0
26.1	Yes	1
48.9	No	0
23.4	No	0

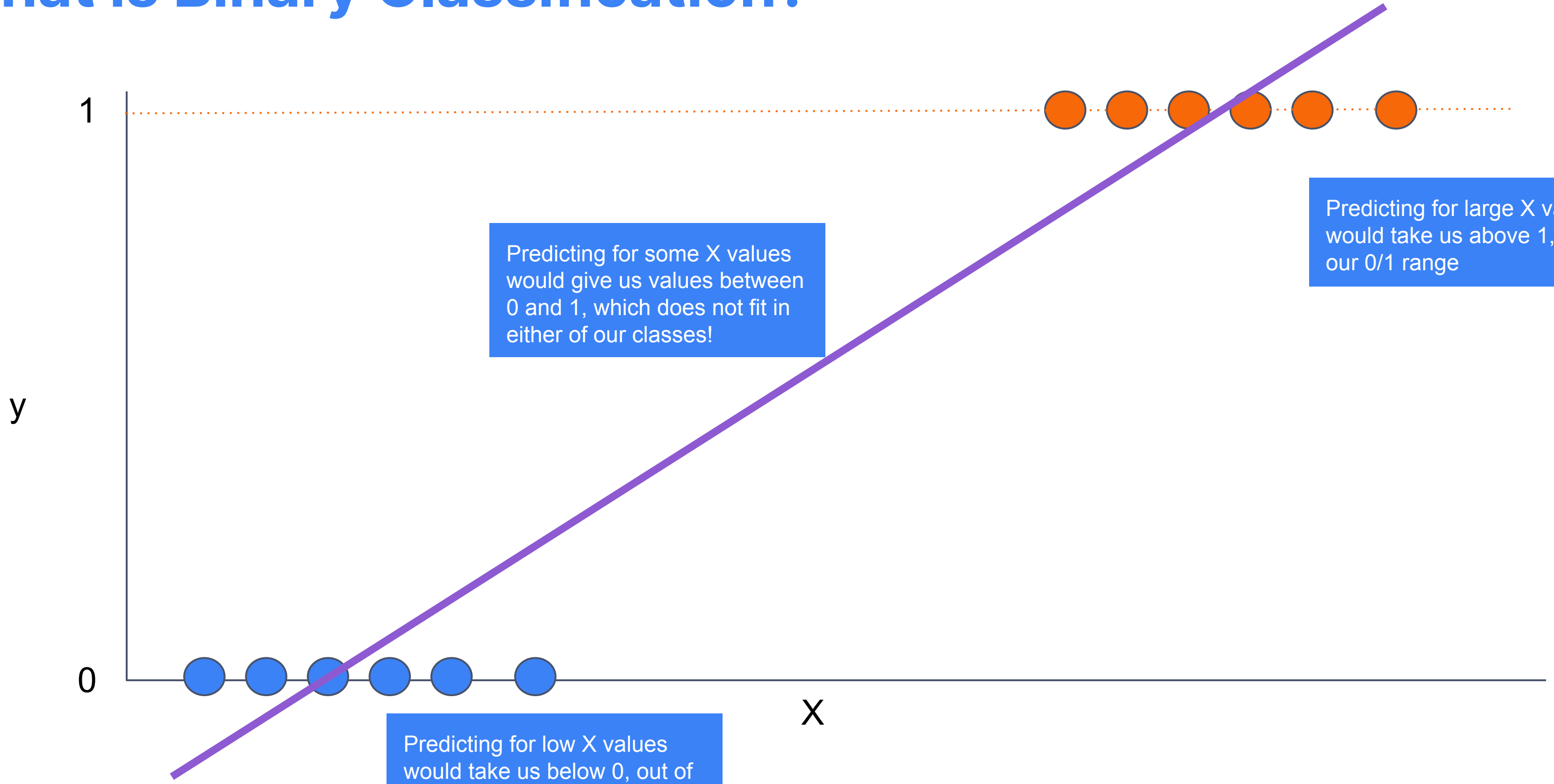
Please do not copy without permission. © ExploreAI 2023.

What is Binary Classification?



Please do not copy without permission. © ExploreAI 2023.

What is Binary Classification?



Predicting for some X values would give us values between 0 and 1, which does not fit in either of our classes!

Predicting for large X values would take us above 1, out of our 0/1 range

Predicting for low X values would take us below 0, out of our 0/1 range

Please do not copy without permission. © ExploreAI 2023.

1 / **Objectives**

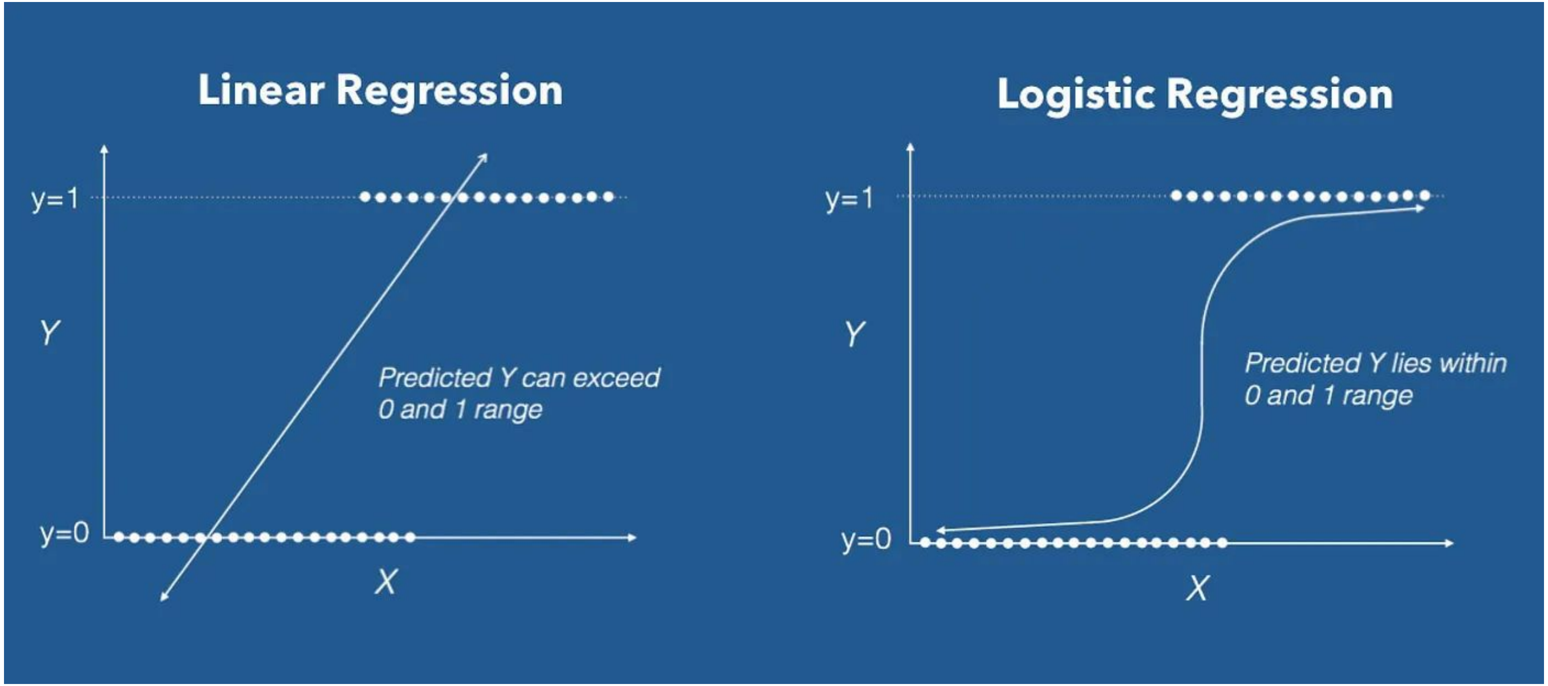
2 / **Binary Classification**

3 / **Logistic Regression**

4 / **To the code!**

What is Logistic Regression?

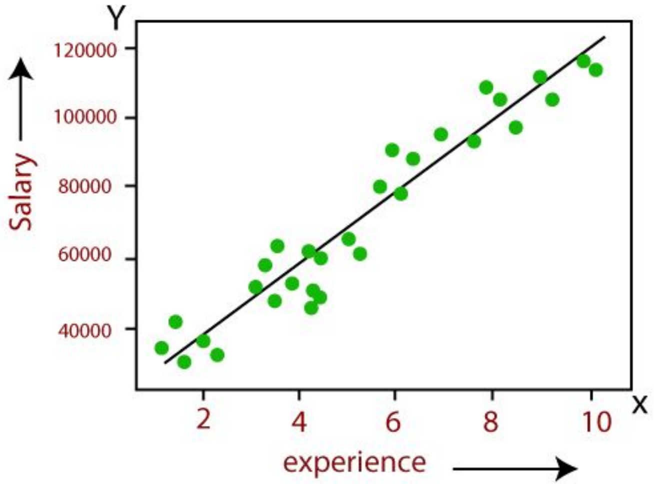
- The answer to our problems (for now)!

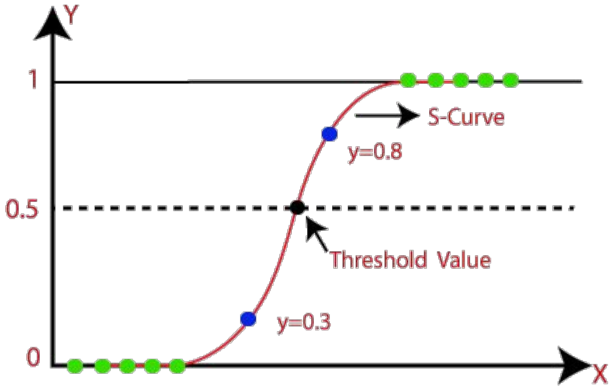

Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed 0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within 0 and 1 range

Please do not copy without permission. © ExploreAI 2023.
Source: https://medium.com/analytics-vidhya/logistic-regression-37955f041b27

What is Logistic Regression?

- Uses a sigmoid curve to squeeze the output values between 0 and 1 (S-shaped curve)
- Why is it still called regression though?
  - Technically, we are still doing a linear regression, we're just squeezing it between 0 and 1
  - Linear regression gives a continuous value of Y as the output
  - Logistic regression gives a continuous value of P(Y=1) as the output

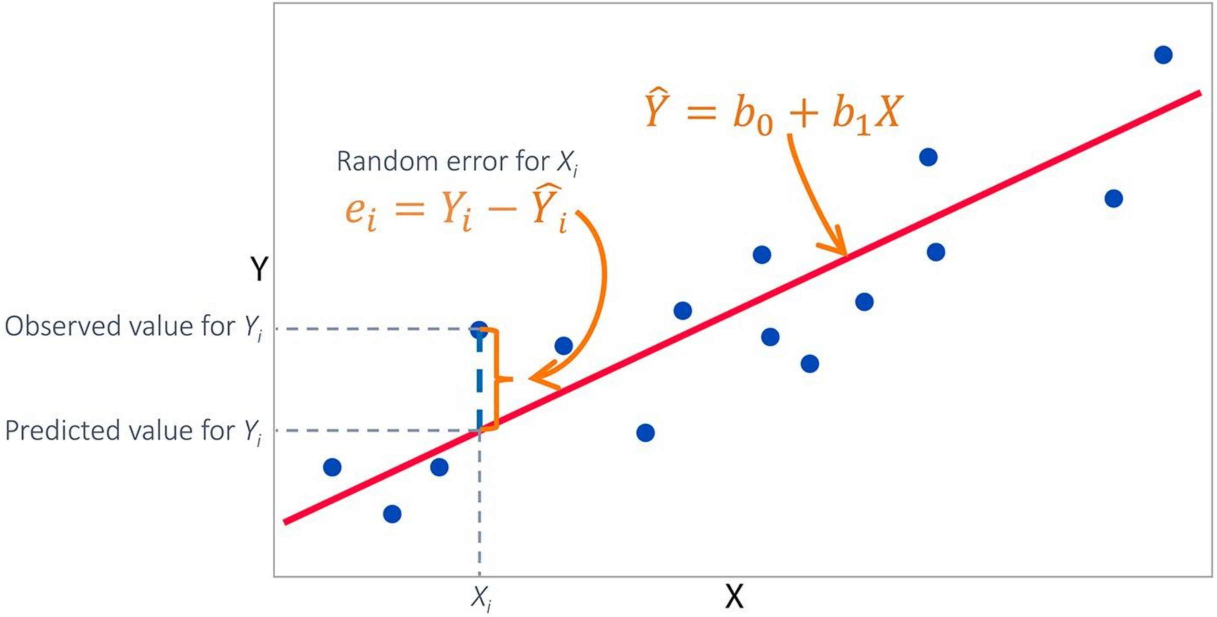




Please do not copy without permission. © ExploreAI 2023.
Source: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

Line of Best Fit

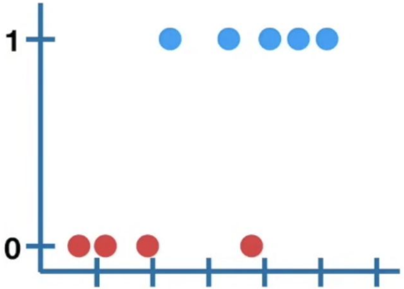
- A throwback to **residuals** from Regression and finding the **line of best fit**...
- We used residuals and **least squares** to fit this line!



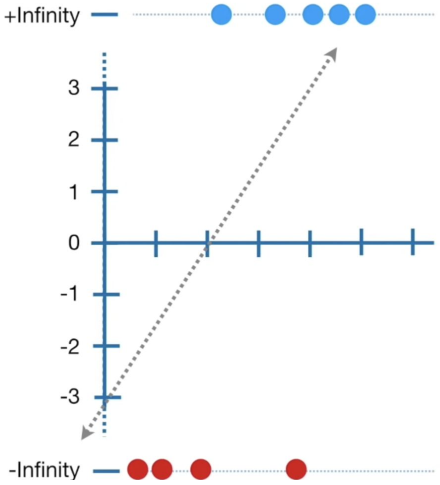
Please do not copy without permission. © ExploreAI 2023.

Maximum Likelihood Estimation

- But how do we find the curve that fits the best for Logistic Regression?
- **Maximum Likelihood Estimation**
  - Parameters are chosen to maximize the likelihood that the assumed model results in the observed data.
- Data is mapped onto a set of axes using the log(odds) on the y-axis



$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$



Please do not copy without permission. © ExploreAI 2023.

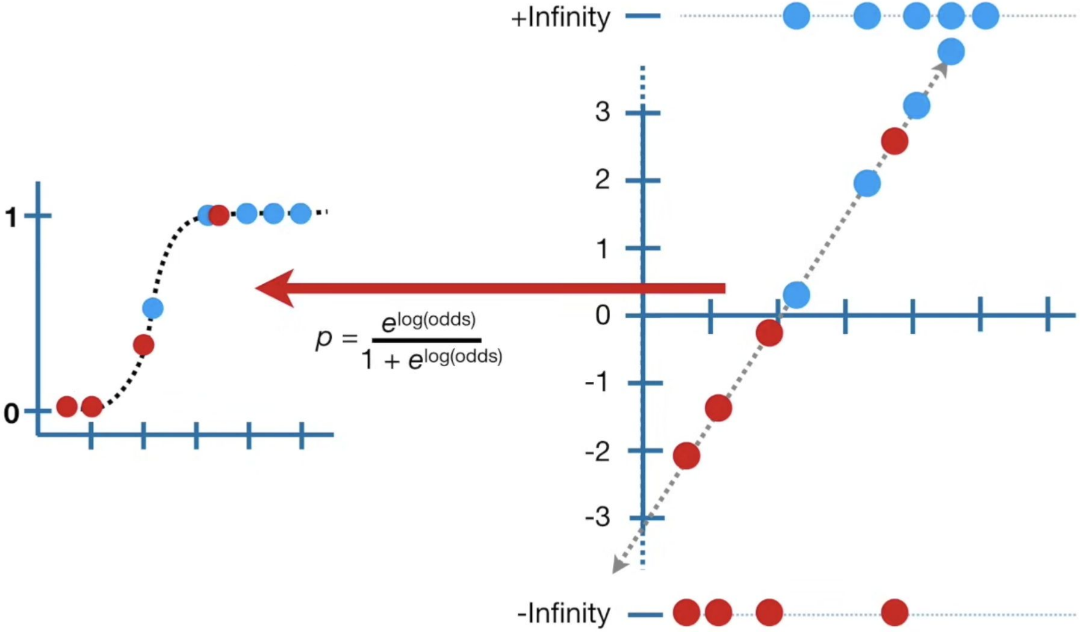
Source: https://www.youtube.com/watch?v=BfKanl1aSG0&t=166s

Maximum Likelihood Estimation

- Each point is mapped onto this line
- We use an equation to map each point back to the original graph based on its log(odds)

Likelihoods of all of these plotted points are then calculated and multiplied together

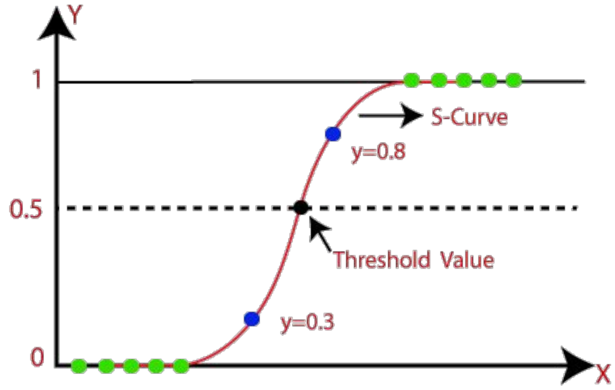
We keep changing the straight line on the log(odds) graph and repeating the process to get the line with the Maximum Likelihood Estimation (i.e. the largest likelihood from the above calculation)



$$p = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

Please do not copy without permission. © ExploreAI 2023.
Source: https://www.youtube.com/watch?v=BfKanl1aSG0&t=166s

Logistic Regression

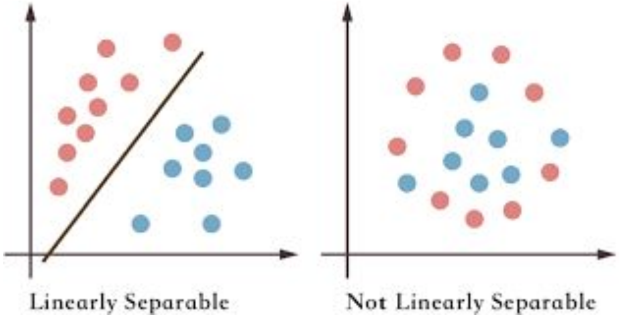


Advantages

- Convenient probability scores for observations
- Collinearity is not as bad as it is for linear regression

Disadvantages

- Can overfit when data is unbalanced (when we have far more observations in one class than the other)
- Doesn't handle large number of categorical variables well.
- Does not work well with **non-linearly separable data**



Please do not copy without permission. © ExploreAI 2023.

Source: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

1 / **Objectives**

2 / **Binary Classification**

3 / **Logistic Regression**

4 / **To the code!**

Questions?

Please do not copy without permission. © ExploreAI 2023.