Welcome to Data Science Online Bootcamp

Week#3_Day#7

dφ

Democratizing Data Science Learning

Learning Objectives

Data Splitting

Logistic Regression

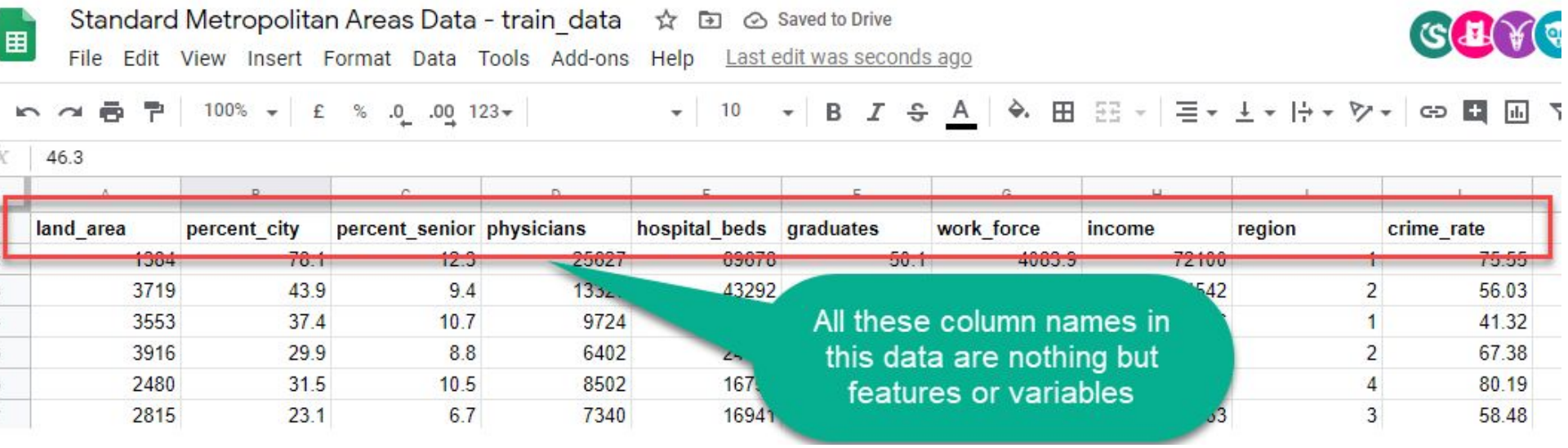
Quick recap of some keywords!

Back to basics ;)

Variables/features

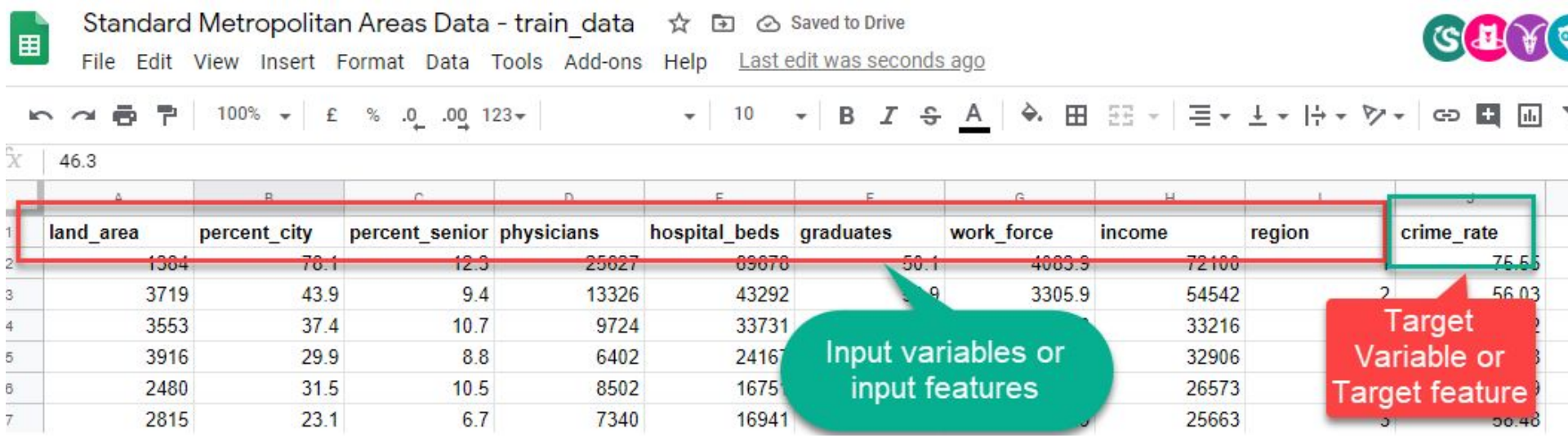
- **Features or Variables:** These are the the most common terms that we would come across from now on.

- **Features and Variables both are the same in a dataset**, they are often interchangeably used. So there is no need to worry about it!

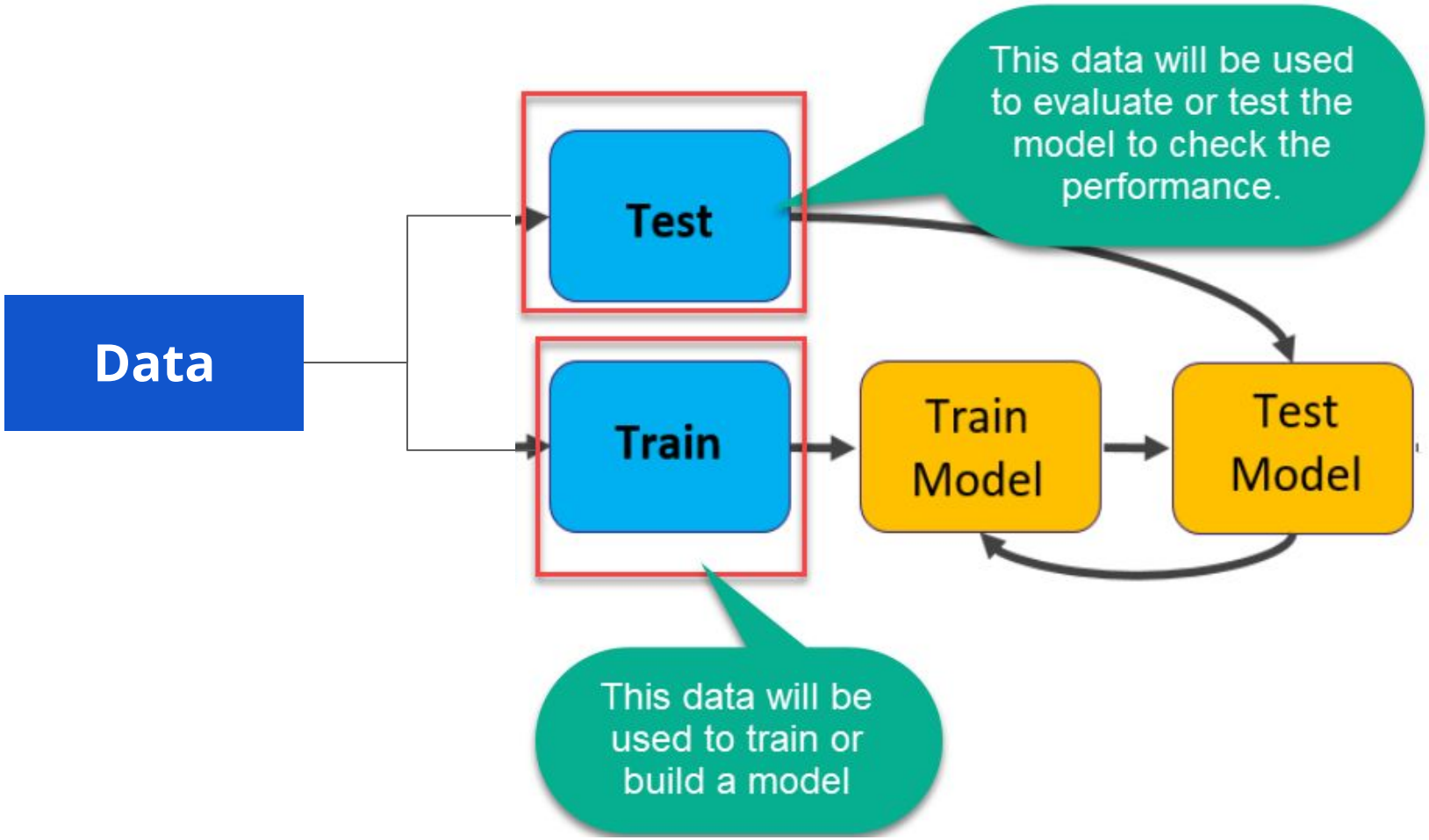


Target and Input variables

- Remember the Standard Metropolitan Areas Data used in previous slides? In that dataset **we might be curious to predict "crime_rate" in future**, so that becomes our target variable and rest of the variables become input variables for building a machine learning model.



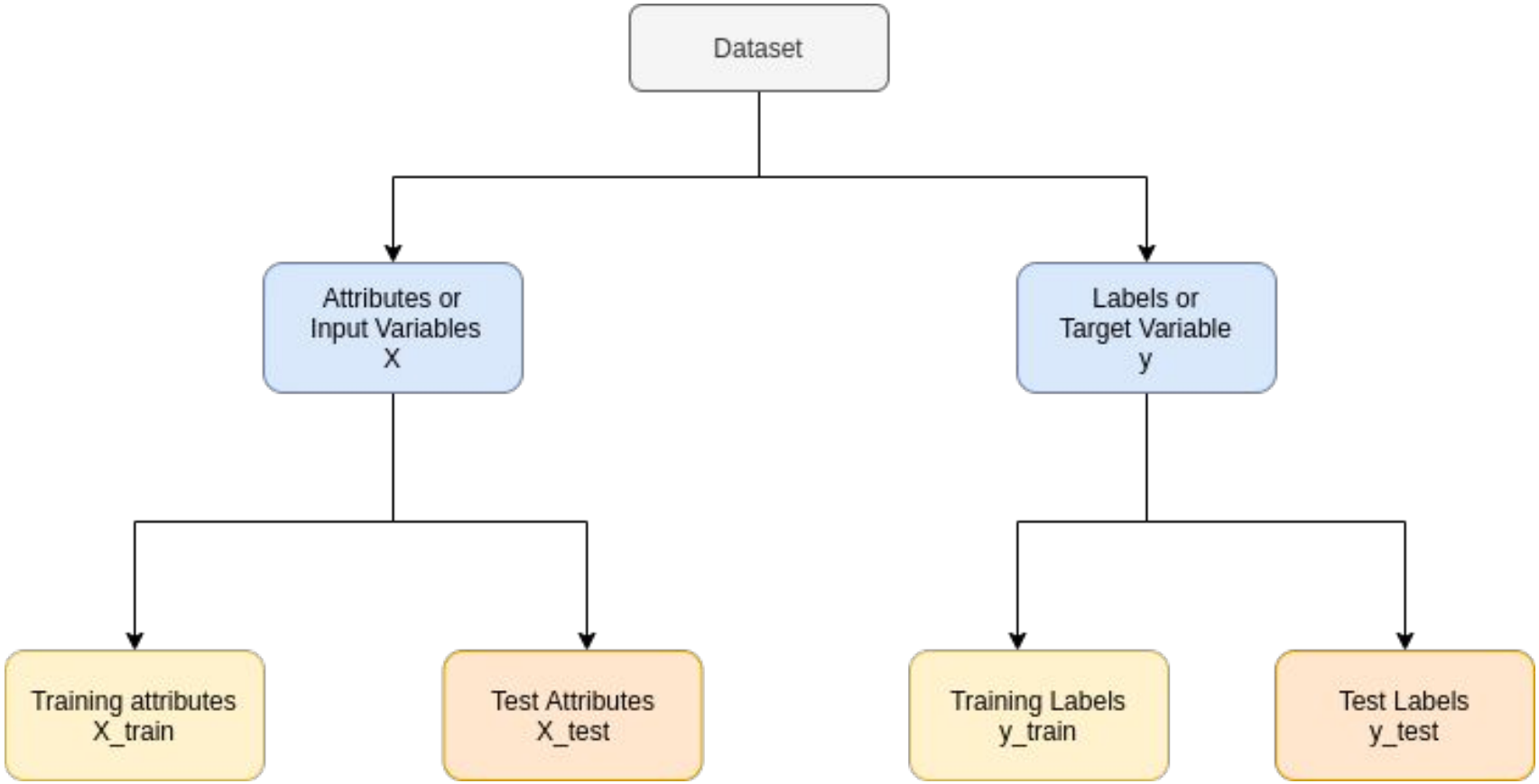
Train and Test Set



We realized some of you are having a hard time with train and test dataset split. Please refer to the image in the next slide

Data Splitting

1. We **separate the data into input and target variable**
2. We further **divide them into train and test datasets**



Data Splitting

Let's consider the following data with 10 entries and our objective is to predict whether someone has purchased a product or not i.e "ProductPurchase" column = Yes or No

Country	Age	Salary	ProductPurchase
Nepal	48	62000	Yes
China	26	38000	Yes
India	30	54000	No
China	28	50000	Yes
India	40		No
Nepal	31	48000	Yes
China		32000	Yes
Nepal	42	89000	Yes
India	53	63000	No
Nepal	36	47000	No

Here, **Country, Age and Salary are the Input variables** used to predict the **target variable - ProductPurchase**.

Separate Input and Target Variables

- We separate the dataset into two new datasets one with input variables - X (Country, Age and Salary) and the other with target variable - y (ProductPurchase).

Adding a snippet from that we used in our first model building exercise.

Separating input variables (X) and target variable (y)

Y has the labels, our answers column. X is all the rest of the data - the features, without the labels (The survived column). This sepration would hoepfully be clearer in a few cells

```
X = titanic.drop('Survived', axis = 1)
y = titanic['Survived']
```

As explained above, X now has the train data without the "Survived" column (this is acheived with the "drop" function). Y, on the other hand, has only the "Survived" column.

Splitting into train and test datasets!

- We'll now split the x dataset into two separate sets — X_train and X_test.
- Similarly, we'll split the dataset y into two sets as well — y_train and y_test.

```
: from sklearn.model_selection import train_test_split
  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

Reference here:
https://github.com/dphi-official/First_ML_Model/blob/master/ML_with_titanic_data.ipynb

Logistic Regression!

Thing to note!

Never jump the gun!

We can take baby steps and learnt things over time!

What's next?

We learnt to build our first machine learning and also learnt about the 2 other ML models (Linear and Logistic Regression). This means we are already in the game and we know how to build models. Ofcourse, there will be ifs and buts, we might be confused because of too many terminologies, many models etc. That is totally common!

Now, let's take some baby steps to understand Logistic Regression. We learn the intuition behind the algorithm first and once we understand it well, we ourselves would be curious to learn the maths behind overtime.

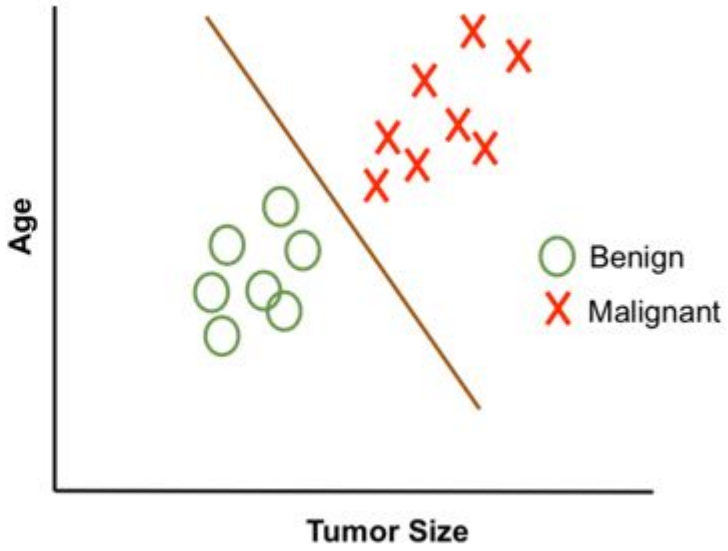
Please note: Getting overwhelmed with maths or coding or tools shouldn't be a blocker for the to learn data science!

What is Classification?

Let's learn with some examples:

- In **Classification we** classify the outcome
- **Examples:**
  - Predict whether a transaction is fraud or not fraud
  - Predict whether to give loan or not
  - Predict whether to give college admission or not
  - Predict the grade (Grade A, B, C, D)
  - Note: Classification can be more than two

Feature	Tumor Age and Tumor Size
Label	Tumor (Benign or Malignant)
Goal/ Aim	We want to predict whether a tumor is benign or malignant from the given age and tumor size



What is Multi-Classification?

It is as simple as dividing waste into 4 categories - plastic, glass, metal, paper

Logistic Regression

- Logistic Regression is one of the basic and popular algorithms to solve a binary classification problems

- For each input, logistic regression outputs a probability that this input belongs to the 2 classes
  - Set a probability threshold boundary and that determines which class the input belongs to

- Binary classification problems (2 classes):
  - Emails (Spam / Not Spam)
  - Credit Card Transactions (Fraudulent / Not Fraudulent)
  - Loan Default (Yes / No)

Logistic Regression

Now, you may ask why don't we use Linear Regression? Why do we need a new algorithm?

Well, you would find all the answers in the video in the next slides.

The video in the next slide is a must watch, the instructor has brilliantly explained about logistic regression!

Must Watch Understanding Logistic Regression

Linear Regression vs Logistic

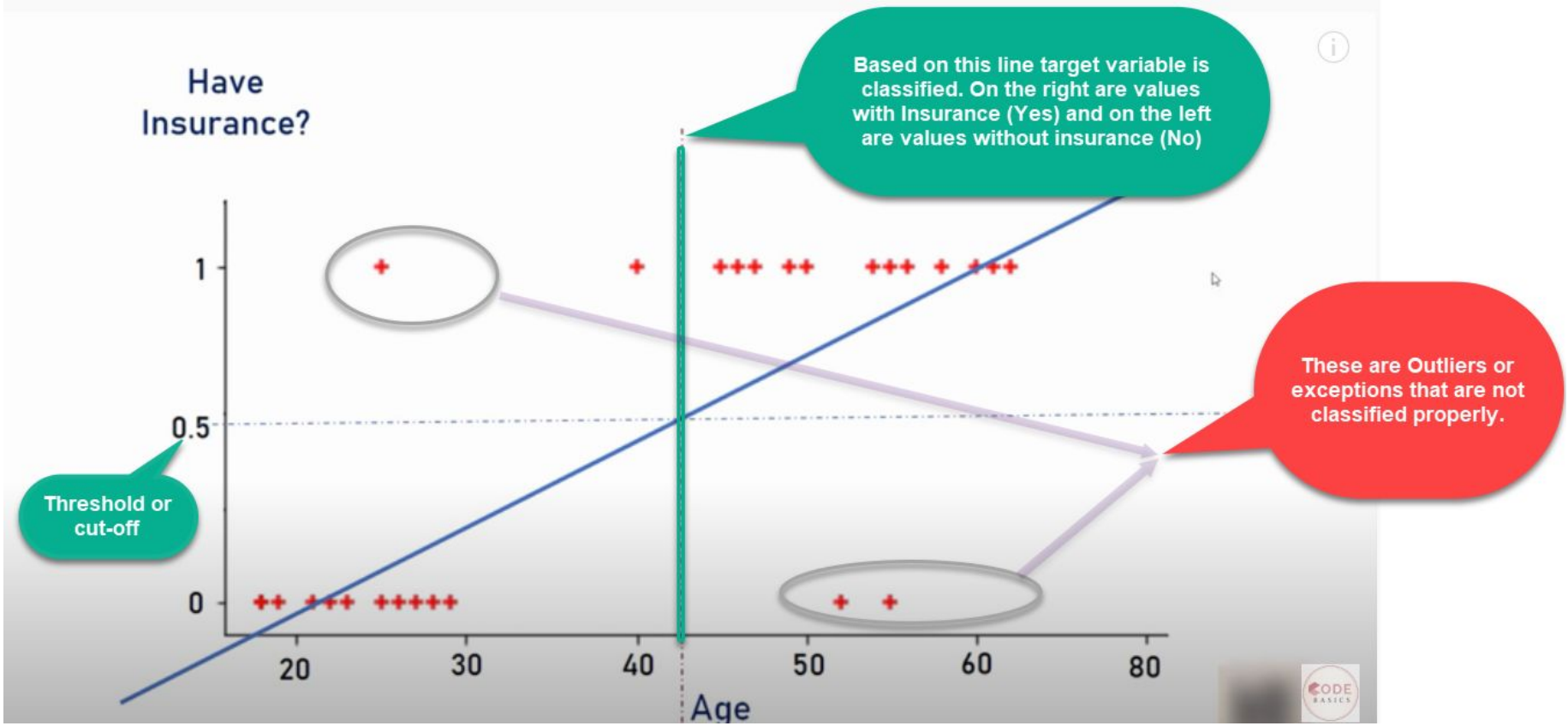
- Linear regression is used to solve regression problems with continuous values
- Logistic regression is used to solve classification problems with discrete categories
  - Binary classification (Classes 0 and 1)
  - Examples:
    - Emails (Spam / Not Spam)
    - Credit Card Transactions (Fraudulent / Not Fraudulent)
    - Loan Default (Yes / No)

Linear Regression vs Logistic

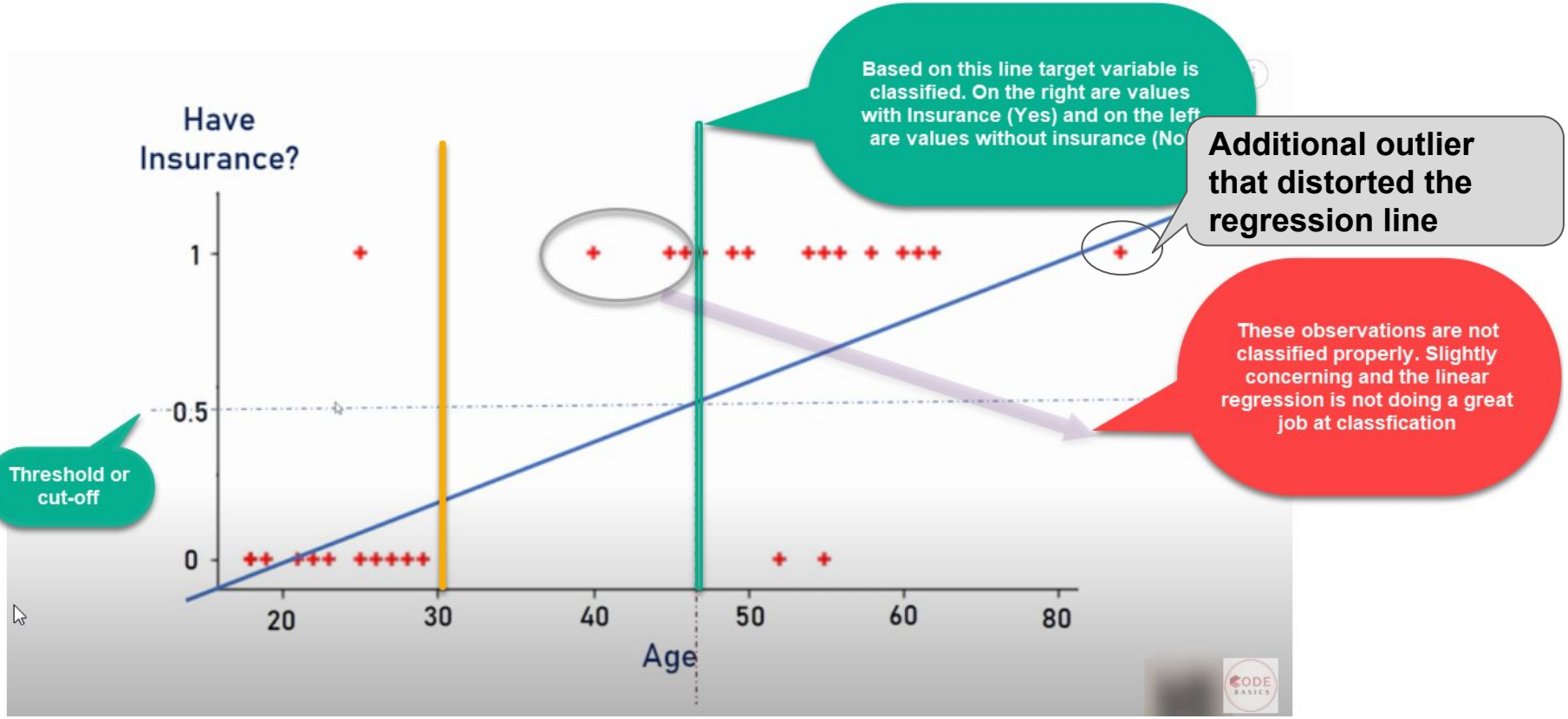
- Let's say a data scientist named John want to predict that whether a customer will buy insurance or not
- Remember that linear regression is used to predict a continuous value where the output (y) may vary between +∞ (positive infinity) to -∞ (negative infinity) whereas in this case, the target variable (y) takes only two discrete values, 0 (No insurance) and 1 (Yes, got the insurance).
- John's decides to extend the concepts of linear regression to fulfil his requirement. One approach is to take the output of linear regression and map it between 0 and 1, if the resultant output is below a certain threshold (say 0.5), classify it as No (didn't buy the insurance) whereas if the resultant output is above a certain threshold, classify it as bought the insurance (yes)

Linear Regression vs Logistic

- We then plot a simple linear regression line and set the threshold as 0.5
  - Negative class (Insurance = No)– Age on the left side
  - Positive class (Insurance = Yes) – Age on the right side



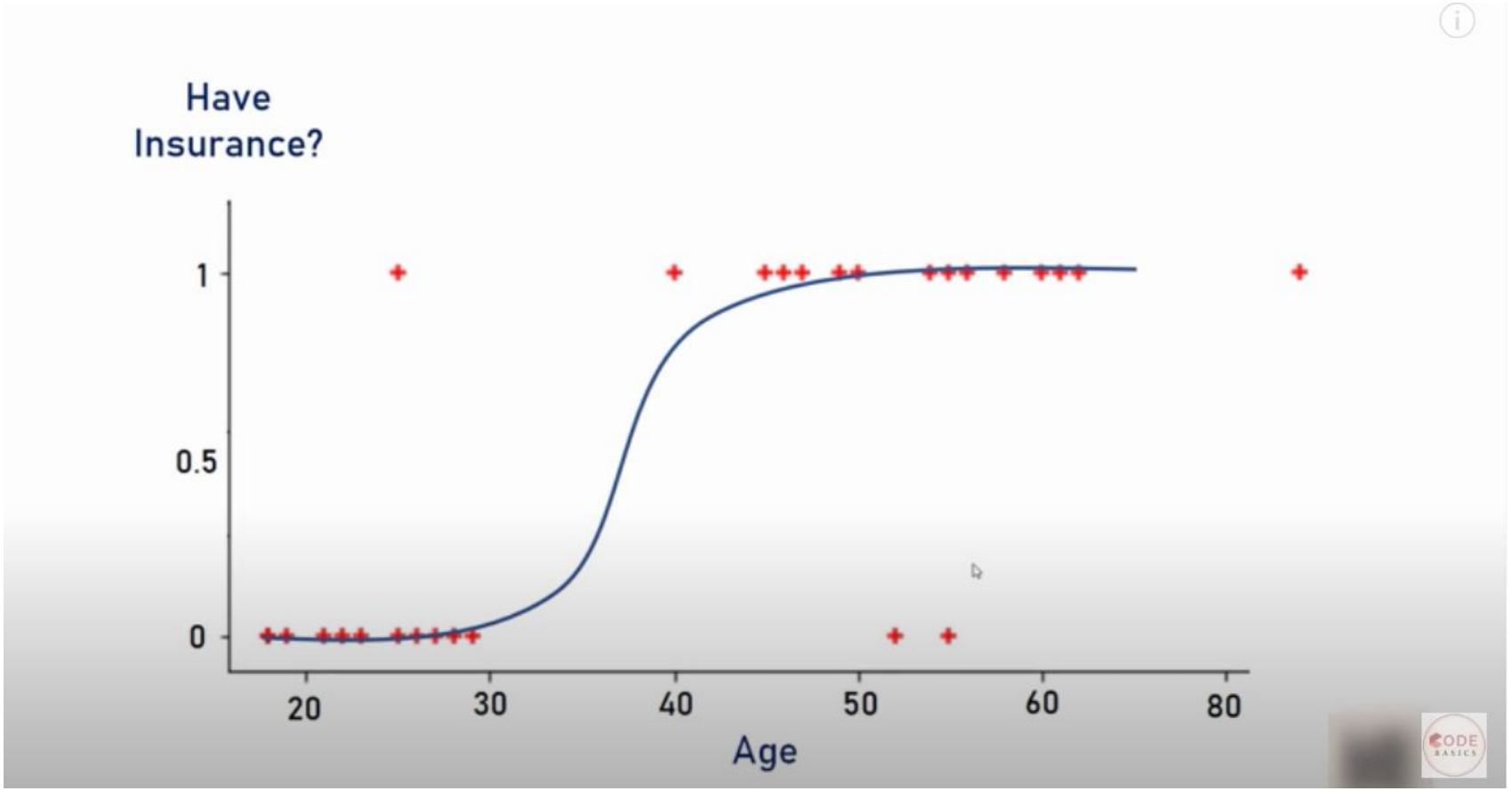
Imagine there is an outlier to towards right



- As we can see outlier in the data and will distort the whole linear regression line.
- Clearly the line is unable to differentiate the classes with the linear line fit
- The line should have been at the vertical yellow line which is able to divide the positive and negative classes i.e yes or no for insurance

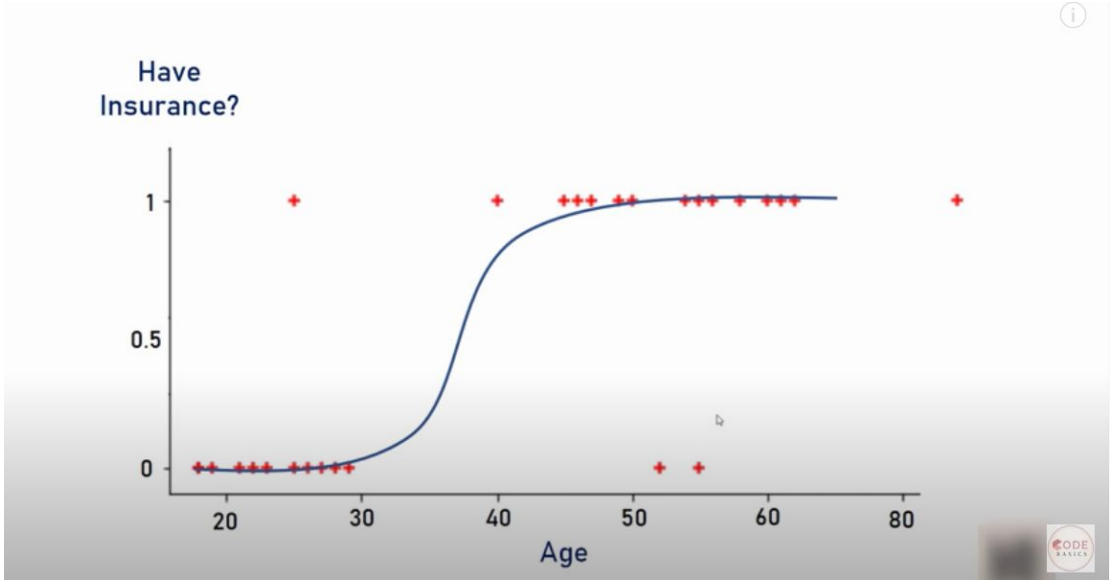
Happy John! (Data Scientist)

- **Well, life would be much simpler if we had a algorithm that would fit the points like below right? It is a much better fit compared to regression line!**



Solution

- Solution – Transform linear regression to a logistic regression curve
- Logistic regression is a Sigmoid function
- Now what does this sigmoid function do?
  - Sigmoid function takes in any real value and gives a output probability between 0 and 1



Logistic vs Linear Regression

- Linear regression is used to solve regression problems with continuous values
- Logistic regression is used to solve classification problems with discrete categories
  - Binary classification (Classes 0 and 1)
  - It can be used for multi-classification problems too

Let’s understand some maths behind it!

It is alright if you don’t understand just try to understand the intuition given in the previous slides

What are we doing in Logistic Regression?

We will use the real-valued output obtained from a linear regression model between 0 and 1 and classify a new example based on a threshold value. The function used to perform this mapping is the **sigmoid function**

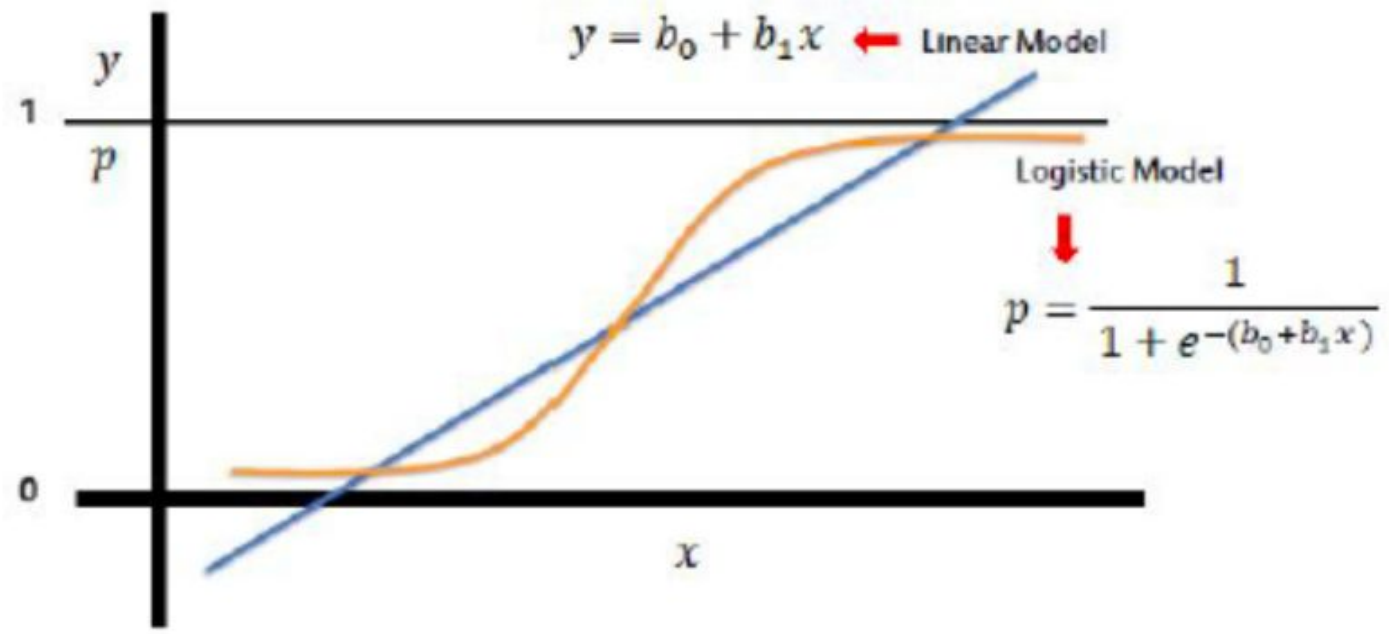
The Sigmoid Function is represented by the formula:

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

There's no need to go into the depth of how we obtained this formula right now.

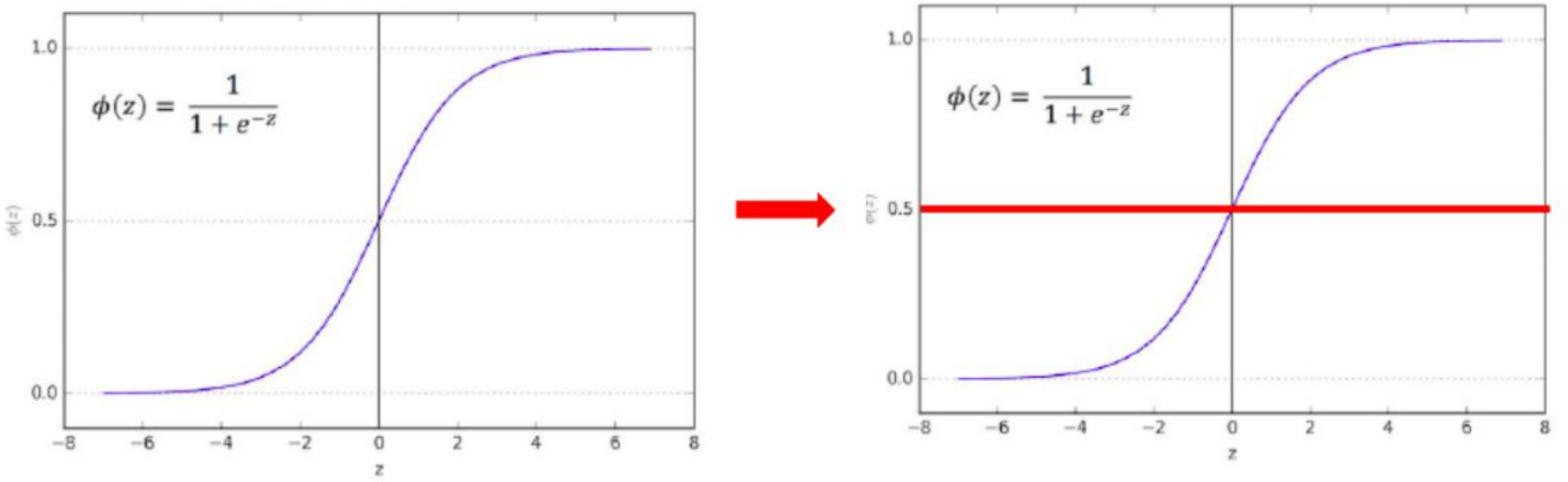
Sigmoid Function (Logistic Function/ Logit)

- Take linear regression function and put it into the Sigmoid function
- Sigmoid function outputs probability between 0 and 1



Sigmoid Function (Logistic Function/ Logit)

- Sigmoid function outputs probability between 0 and 1
- Default probability threshold is set at 0.5 typically
  - Class 0 – Below 0.5
  - Class 1 – Above 0.5



Types of Logistic Regression

Logistic Regression model can be classified into three groups based on the target variable categories:

1. **Binary Logistic Regression**
   - The target variable has two possible categories.
   - Common examples : 0 or 1, yes or no, true or false, spam or no spam, pass or fail, Transactions (Fraudulent / Not Fraudulent), Medical Condition (Diseased/ Not diseased)

2. **Multi- Class Logistic Regression**

   a. **Multinomial Logistic Regression**
      - The target variable has three or more categories which are not in any particular order. So, there are three or more nominal categories.
      - Examples: Fruits (apple, mango, orange and banana), profession (e.g., with five groups: surgeon, doctor, nurse, dentist, therapist)

   b. **Ordinal Logistic Regression**
      - The target variable has three or more ordinal categories. So, there is intrinsic order involved with the categories.
      - For example, the student performance can be categorized as poor, average, good and excellent.

We will be discussing about Multi-Class Logistic Regression in tomorrow's module!

Notebook on Logistic Regression

https://github.com/dphi-official/ML_Models/blob/master/Logistic_Regression/logistic_regression.ipynb

Let's do some practice

Instruction:

Use the raw data github link: https://raw.githubusercontent.com/dphi-official/Datasets/master/HR_comma_sep.csv

Or you can download it here from here

Exercise:

- Load libraries and data.
- Do some exploratory data analysis to figure out which variables have direct and clear impact on employee retention (i.e. whether they leave the company or continue to work)
- Plot bar charts showing impact of employee salaries on retention
- See the correlation between department and employee retention
- Separate dependent and independent variables.
- Split the data into train set and test set
- Now build Logistic Regression model and do prediction for test data
- Measure the accuracy of the model

Slide Download Link

- You can download the slides here: https://docs.google.com/presentation/d/10mkWll_9s8LkZpy-SzRWQjFr2ay6VppgpjkBVs2HSvM/edit?usp=sharing

That’s it for the day. Thank you!

Feel free to post any queries in the #help channel on Slack