LOVELY PROFESSIONAL UNIVERSITY

Report On

Predicting the Titanic Survival Rate

Subject: INT247 (Machine Learning Foundation)

Team:

Roll. No	Reg. No	Name of student
10	11706732	Karnam Abinay Goud
12	11706922	Tanari Bhargav
31	11714590	Putta Bharath

Submitted to:

Dr. Aditya Khamparia (Head of the Dept. Machine Learning)

Introduction:

In machine learning and statistics classification is a problem of identifying to which of a set of categories a new observation belongs to based on the *training data*.

Using logistic regression we can solve classification problems where we are trying to predict discrete values.

The *convention* for binary classification is to have two classes 0 and 1

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.
Goal :In this Project, we will build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although

many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).
In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

Types of Logistic Regression

1. Binary Logistic Regression

The categorical response has only two 2 possible outcomes. Example: Spam or Not

2. Multinomial Logistic Regression

Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

3. Ordinal Logistic Regression

Three or more categories with ordering. Example: Movie rating from 1 to 5

Sigmoid Function

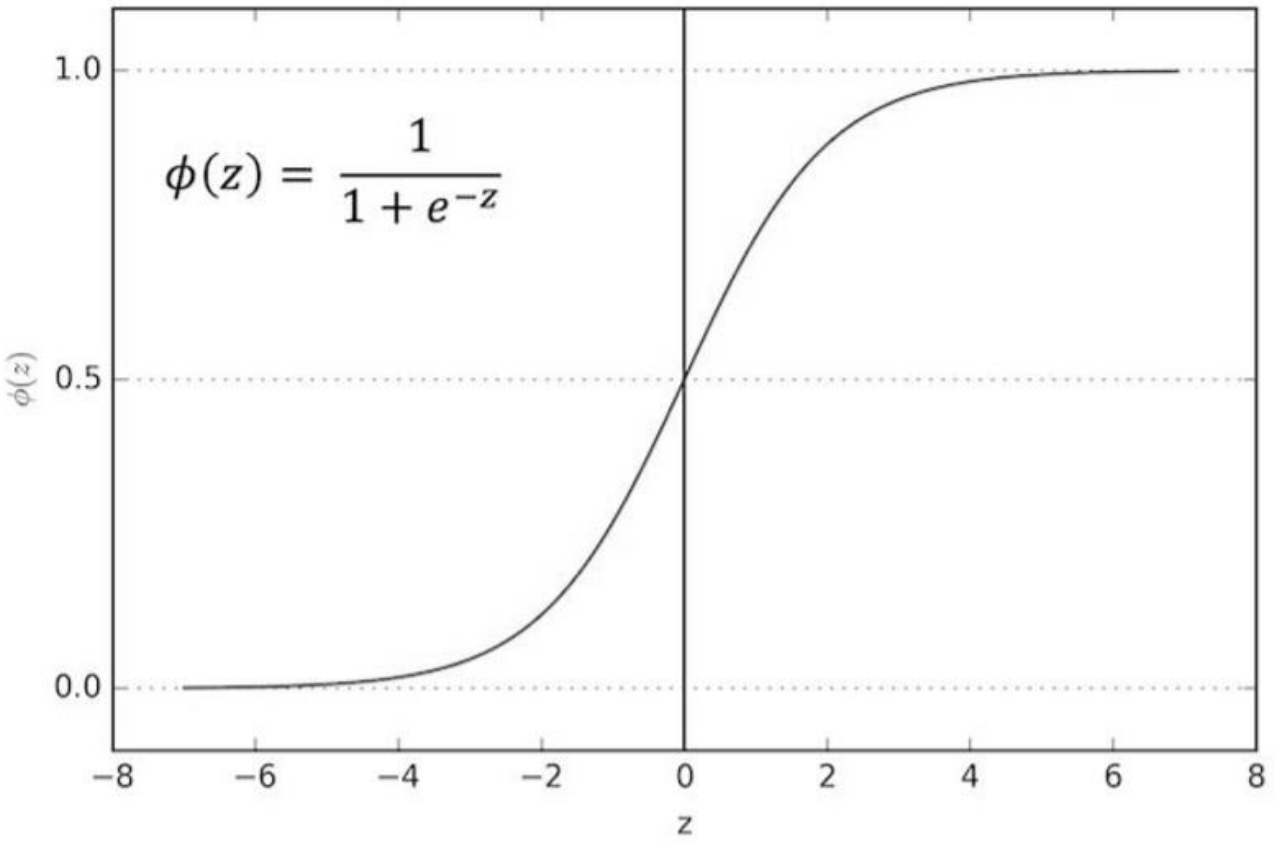
The sigmoid function also known as the logistic function is going to be the key to using logistic regression to perform classification.

- The sigmoid function takes in any value and outputs it to be between 0 and 1.

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

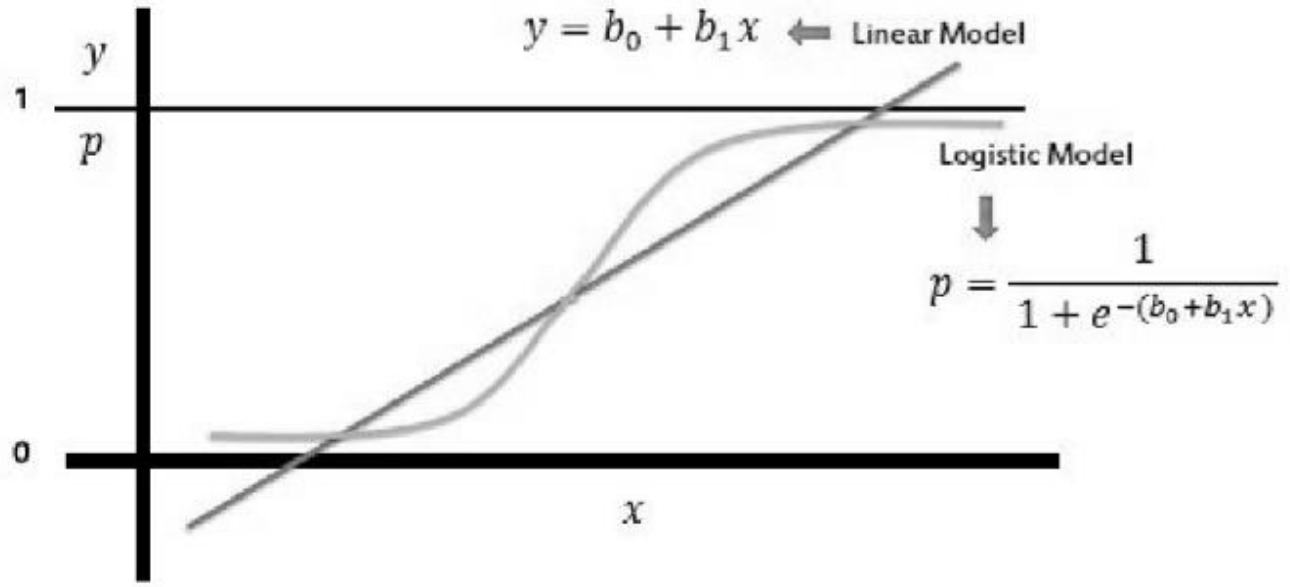
Sigmoid Function

The key thing to notice here is that it doesn't matter what value of **z** you put into the logistics or the sigmoid function you'll always get a value between 0 and 1.



Sigmoid Function Curve

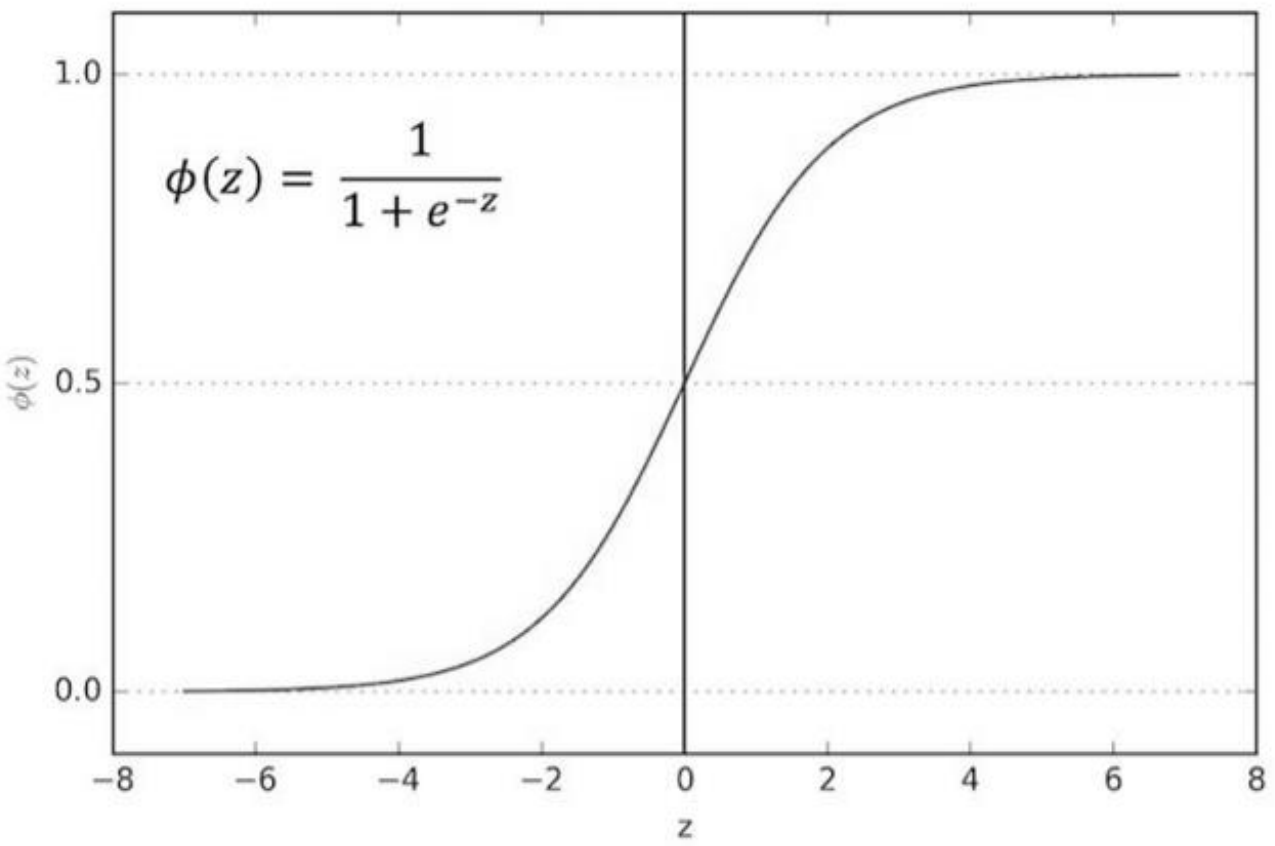
- This means we can take our linear regression solution and place it into the sigmoid function and it looks something like this:



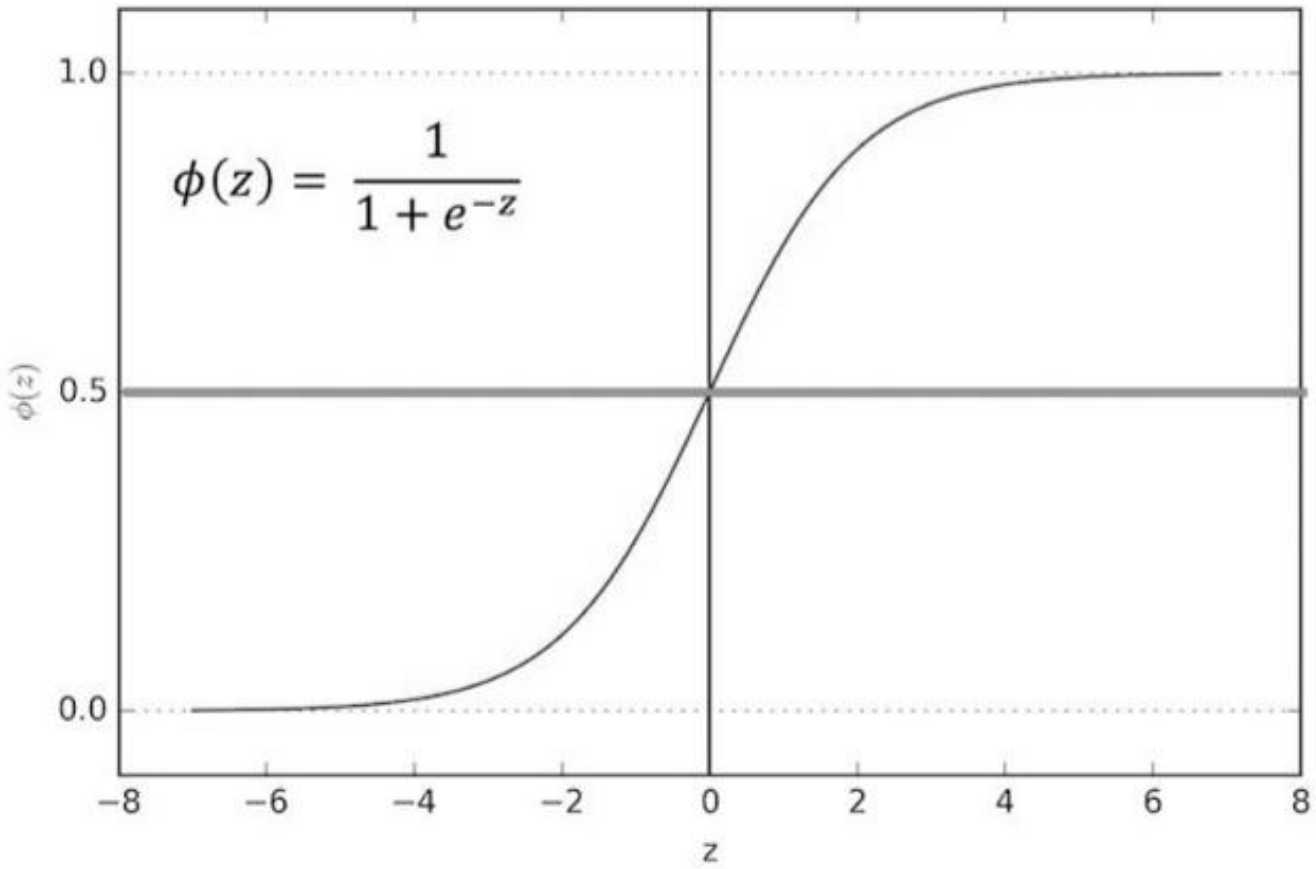
Linear Curve in Logistic Regression Curve

- If you take that linear model and place it into a sigmoid function then we are finally able to *transform* linear regression to logistic model meaning it doesn't matter whatever the value of linear model output actually is it's always going to be between 0 and 1 when you place it into the sigmoid function**.**

This results in a probability from 0 to 1 belonging in class 1.



- We can set a cutoff point at 0.5 and we can say anything below 0.5 results in class 0 and anything above 0.5 belongs to class 1.



Logistic Regression Curve with cut-off point

So we are going to transform that 0.5 probability as a cut off point.

Model evaluation

After we have trained a logistic regression model on some training dataset we can evaluate the model's *performance* on some test dataset, we can use confusion matrix to *evaluate* classification models.

Confusion matrix:

The confusion matrix is a table test is often used to describe the *performance* of the classification model on the *test* data for which the *true* values are already known, so we can use a confusion matrix to evaluate a model.

#example: testing the presence of a disease

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

NO = negative test = False = 0

YES = positive test = True = 1

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Basic Terms:

- **True Positives(TP)**= are the cases in which we *predicted yes* they have the disease and in reality, they *do have* the disease.
- **True Negative(TN)**= are the cases in which we *predicted no* they don't have the disease and in reality, *they don't* have the disease.
- **False Positive(FP)** = are the cases in which we *predicted yes* they have the disease and in reality, they don't have the disease. This is also known as Type 1 Error.
- **False Negative(FN)**= are the cases in which we *predicted no* they don't have the disease and in reality, they do have the disease. This is also known as the Type 2 Error**.**

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Accuracy:

how often is it correct?

Accuracy = (TP+TN)/Total

Accuracy = (100+50)/165 = 0.91

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

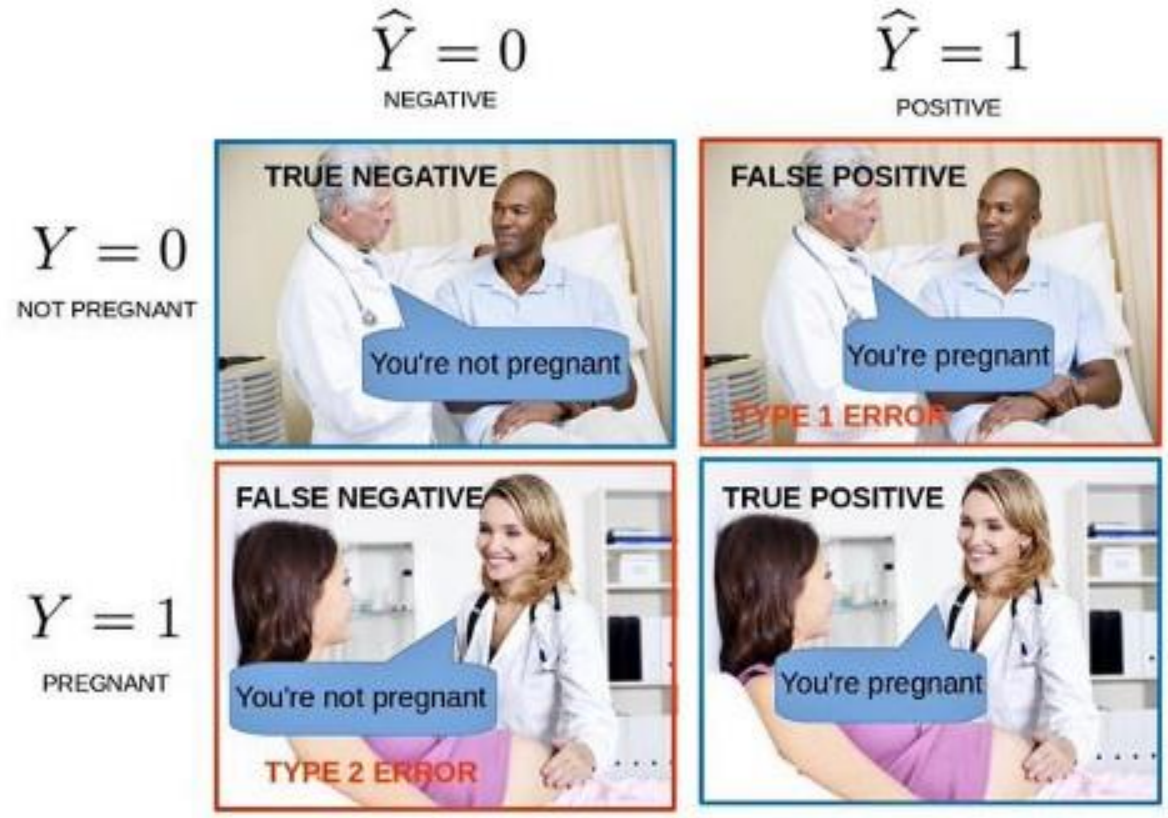
Misclassification Rate:

How often it is wrong ?

MR = (FP+FN)/total

MR = (10+5)/165 = 0.09

This is also called as the Error Rate.



Type 1 and Type 2 error

Type of Errors:

1. Type 1 Error (False Positive)
2. Type 2 Error (False Negative)

Advantages:

- it doesn’t require high computational power
- is easily interpretable
- is used widely by the data analyst and data scientists.
- is very easy to implement
- it doesn’t require scaling of features
- it provides a probability score *for* observations.

Disadvantages:

- while working with Logistic regression you are not able to handle a large number of categorical features/variables.
- it is vulnerable to overfitting
- it can’t solve the non-linear problem with the logistic regression *model* that is why it requires a transformation of non-linear features
- Logistic regression will not perform well with independent*(*X*)* variables that are not correlated to the target*(Y)* variable.