Logistic Regression

BY MG ANALYTICS

What are we trying to predict?

- Classes and not numbers, cat-Dog, Pass-Fail, Approval-Rejection.
- Two classes which are not possible together like someone cannot pass and fail an exam at the same time.
- Probability of outcome being in favour of one class as finding out just hard class may not be enough to clarify understanding.
- e.g. predicting whether somebody might have a heart disease given their age. Predicting "Yes" for both Age=45 and Age=85 is not informative enough.
- Probability of outcome being "Yes" is much more informative and lets us differentiate between different predictor values and their effect
- That implies , we will be trying to predict $P(Y=1 | X)$

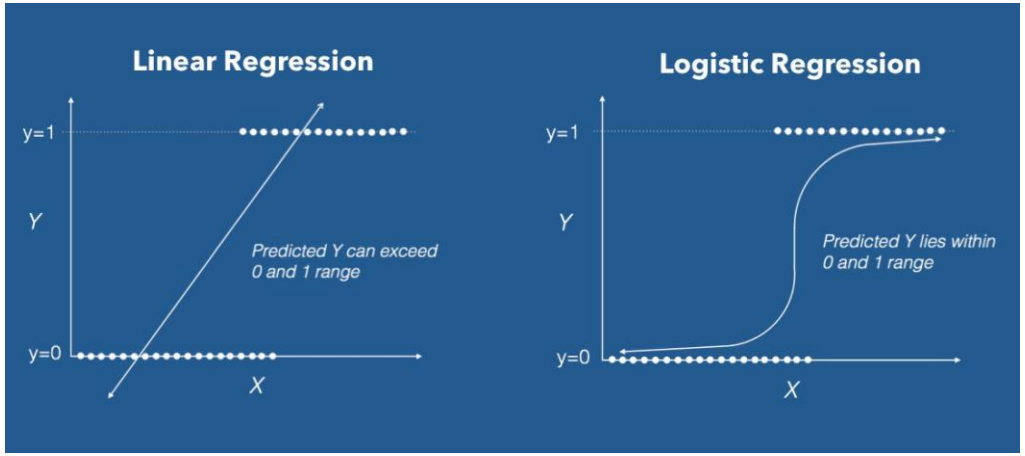
Linear Regression vs Logistic Regression

MG ANALYTICS

Criteria	Linear Regression	Logistic Regression
Target Values	continuous dependent variable.	Hard classes 0 and 1 assessed from probability values which range from 0-1
Attributes / Features	independent variables can be continuous or discrete.	independent variables can be continuous or discrete
Equation Formed	$Y = b_0 + b_1 \times X + e$	$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x_{1i} + \beta_2 * x_{2i} \ldots.)}}$
Solves Problems Like prediction of	Temperature, marks of student, future sales	Binary classes like cat-Dog, Pass-Fail, Approval-Rejection

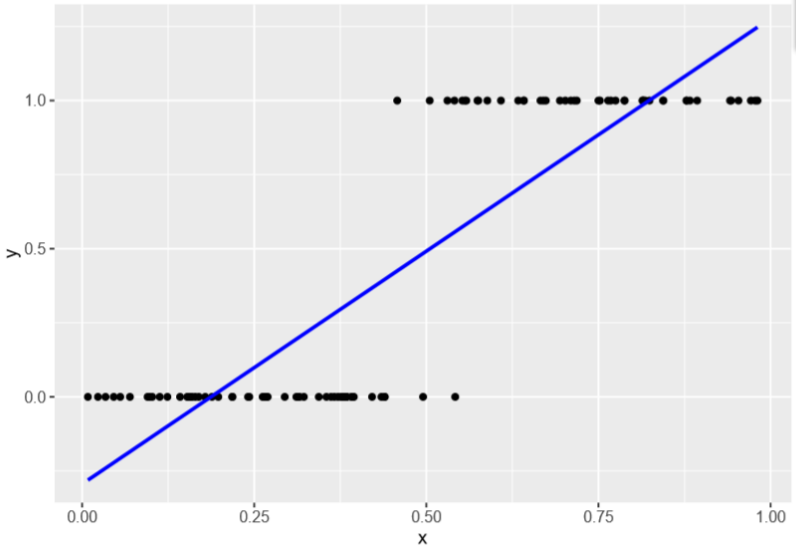
Definition

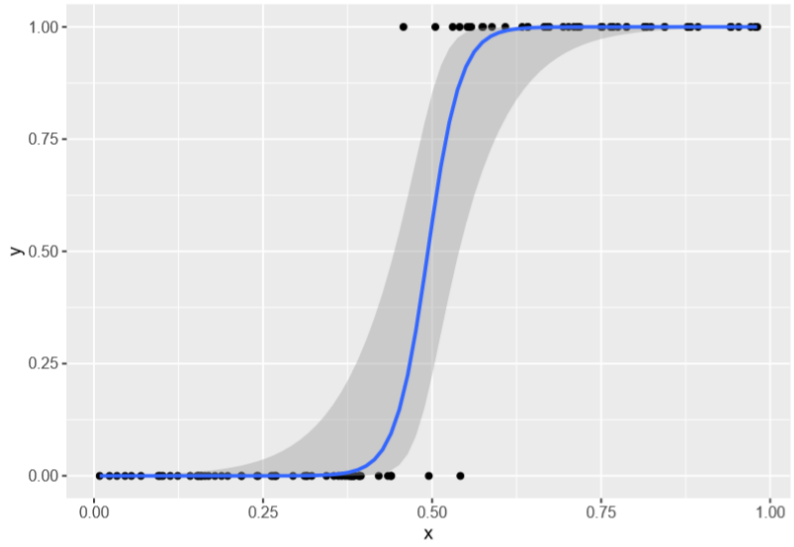
- The **logistic regression** technique is used for finding out results which can be represented in the binary (0 or 1, true or false, yes or no) values, means that the outcome could only be in either one form of two. For example, it can be utilized when we need to find the probability of success or failure of event.



Objective

- Linear equation give result as a continuous value which would not be of much use if we want to classify target into two classes.
- To map linear regression equation such that we could get a curve as shown in figure2. which could give us values close to 0 and 1 for given attribute X.





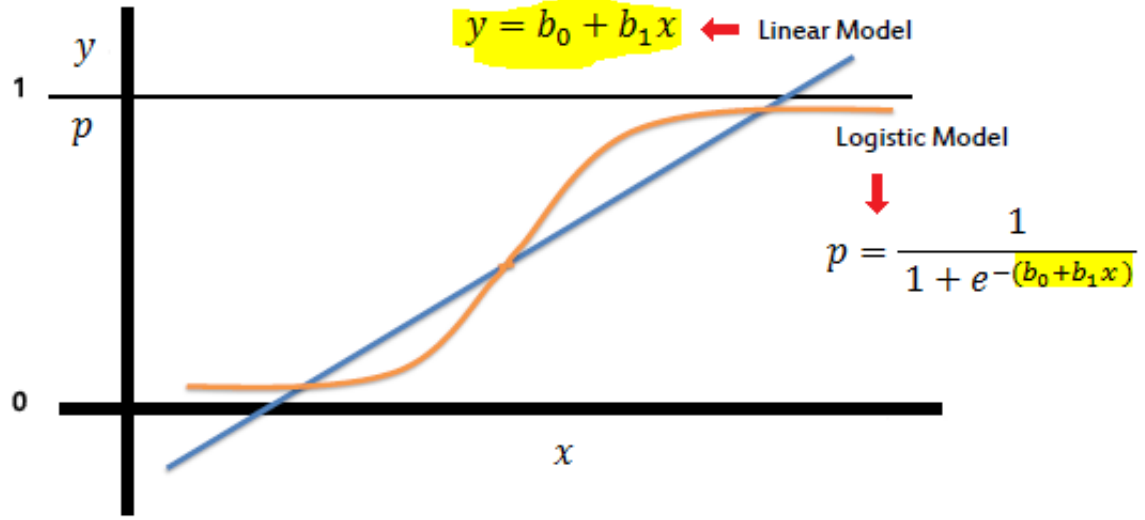
Why is it a "regression" if it finds out Binary classes?

- it's underlying technique is quite the same as Linear Regression. The term "Logistic" is taken from the **Logit function** that is used in this method of classification.

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

$$t = \beta_0 + \beta_1 x$$

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



Getting intuition via example

Age	Response
4	yes
4	yes
4	yes
4	yes
4	no
5	yes
5	yes
5	yes
5	no
5	no
6	yes
6	yes
6	no
6	no
6	no
7	yes
7	no
7	no
7	no
7	no

MG ANALYTICS

- if someone asked you what might the response be if the child's age is 7.
- By looking at the table your guess would be "no". What you did there was to look at probability of response being "no" when age is 7. And you naturally guessed for "no" because that had higher probability [chances]
- So instead of modeling y we should model P(y="yes") or P(y="no"). lets denote that by just p.
  - Instead of **$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 .... \beta p * xp$** .
  - we'd model this: **$p = \beta 0 + \beta 1 * x1 + \beta 2 * x2 .... \beta p * xp$**
  - but this is problematic because right hand side in equation above can take values in the interval **$(-\infty, +\infty)$** .where as probability p can take values in **[0,1].** We need to transform this so that ranges match on both sides.
- **$\log(p/(1-p)) = \beta 0 + \beta 1 * x1 + \beta 2 * x2 .... \beta p * xp$** : This takes care of the range mismatch issue.

Odds and log(odds)

- Odds in favour of a team winning a game is 1 is to 4, where 5 is total number of games.
- Written as ¼
- Probability is 1/5 in above case.
- Odds are not probability, it is ratio of something happening over something not happening
- While Probability is odds of something happening over all the possible cases.
- O(W)= 5/3 = 1.7
- P(W)= 5/8 = 0.625, P(L)= 3/8= 0.375 or P(L)=1-P(W) = 1-0.625 =0.375
- P(W)/P(L)= P(W)/1-P(W)=(5/8)/(3/8)=5/3=O(W) → p/1-p
- Odds range from 0-1 for against scenario while range from 1 to infinity for favourable case.

$$P(y_i = 1|X_i) => p_i$$

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 * x_{1i} + \beta_2 * x_{2i}....$$

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x_{1i} + \beta_2 * x_{2i}....)}}$$

On solving…

What do we do with this equation?

The objective here is not to "correctly" estimate log(p/(1−p))

We want to either maximize or minimize value of p such that we get either a low probability or high probability i.e. close to 0 or 1.

Likelihood

- Lets say P(yi =1 | xi) is Probability of y being1 given x
- whenever y =1 : L = P and
  when y =0 : L =1−P .
- when y =1, Li equals to probability of y being 1, when y =0, L equals to probability of y being 0.
- We want our probabilities to match with real outcome. Hence we would like to maximize L.

Likelihood for one Observation

$$p_i^{y_i} * (1 - p_i)^{(1-y_i)}$$

Likelihood for all observation

$$\prod_{i=1}^{n} p_i^{y_i} * (1 - p_i)^{(1-y_i)}$$

Cost Function

$$-\sum_{i=1}^{n} y_i * log(p_i) + (1 - y_i) * log(1 - p_i)$$

Confusion Matrix

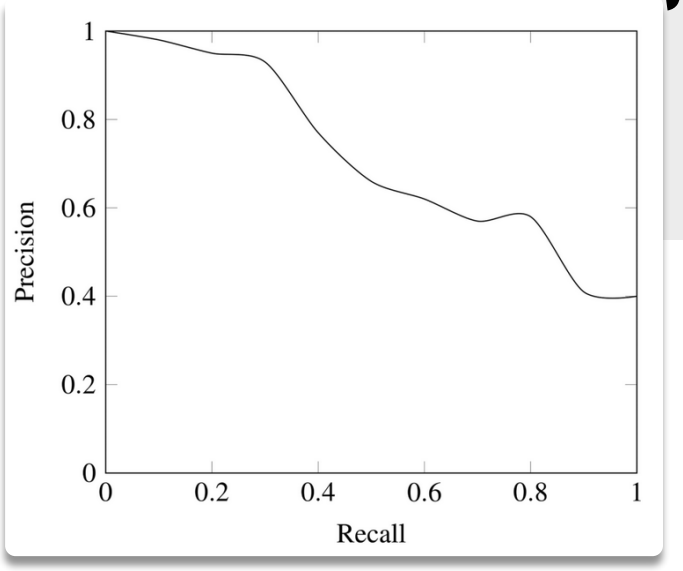
$$p_i > cutoff => y_i = 1$$
$$p_i <= cutoff => y_i = 0$$

Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

- **ACCURACY:** Correctly predicted values out of all values.
  - (TP + TN) / (P + N)
- **PRECISION:** Correctly predicted positives out of all positive predicted values.
  - TP / (TP + FP)
- **SENSITIVITY / RECALL:** out of all the total actual positive, how many positive are predicted correctly. Also known as **TPR**.
  - TP / TP+FN
- **SPECIFICITY:** Correctly predicted negatives out of all negatives.
  - TN / FP+TN

What metric to use?

- We cannot decide upon a single metric hence we try to use combination of metrics like Precision and recall.
  - ***Precision = TP / (TP + FP)***
  - ***Recall = TP / (TP + FN)***
- Example: Predicting Cancer amongst Patients only if we are absolutely sure means having a high cut off value, in this case we might miss out on a few actual positives.
  - Very Few FP **>->** High Precision, low recall.
- Example: Predicting Cancer amongst Patients by avoiding missing any positive cases means having a low cut off value, in this case we want to avoid any false negatives.
  - Very few FN **>->** High Recall, low Precision.
- Hence as we lower the cut off value Recall increases and Precision decreases



MG ANALYTICS

- **type 1 error or FPR** = FP / FP + TN
- If data is **Balanced** than we will calculate accuracy rate and formula for that is :
  - Accuracy : TP + TN / TP +FP + FN +TN
- For **imbalanced** dataset we use recall , precision and F1
- If FP is important go to precision and if FN is important go to recall
- **F beta** score = (1+beta)*2 *precision * recall/ (beta*2* recall) + precision
- If both FN and FP are important than beta is 1
- If FP is important than we decrease beta values
- If FN is important than we increase beta values

Should we average Precision-Recall?

	Precision	Recall	average	F1 score
Model 1	0.5	0.4	0.45	0.444
Model 2	0.9	0.1	0.5	0.18
Model 3	0.02	1.0	0.51	0.0392

Average = P + R /2

F1Score = 2(P*R)/(P+R)

Predict y=1 all the time

Miss out on lots of
Positive cases.

MG ANALYTICS

AUC Score

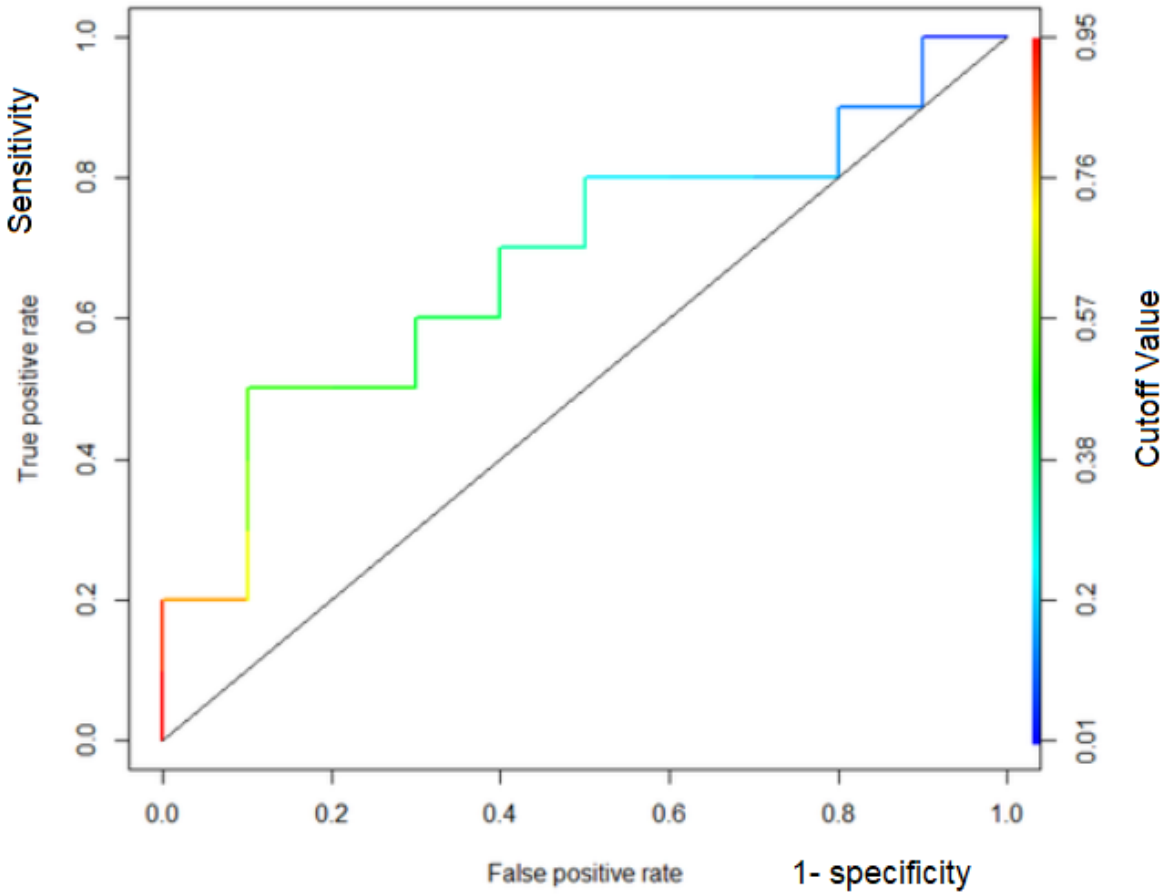
- AUC indicates how well the probabilities from the positive classes are separated from the negative classes.
- It tells how much model is capable of distinguishing between classes.
- Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.
- Sensitivity = "***True Positive Rate***"
- (1 — Specificity) = "***False Positive Rate***".

$$Specificity = \frac{TN}{TN + FP}$$

$$1 - Specificity = 1 - \frac{TN}{TN + FP}$$

$$1 - Specificity = \frac{TN + FP - TN}{TN + FP}$$

$$1 - Specificity = \frac{FP}{TN + FP}$$



- Its used in binary classification problem
- Supposed we are implementing logistic regression and suppose the model has predicted some probabilities then we need decide what is threshold or cutoff
- Threshold value or cutoff is decided by problem statement.
- For example : if we need higher FPR or lower TPR based upon that we can play with cutoff

For constructing roc and auc curve we need to consider few threshold values

- For ROC we require both FPR and TPR
- Based upon the threshold values we get TPR and FPR
- It plots a graph , joining which point we get a curve and area under curve called AUC curve
- Higher AUC the better the model is
- Good model is greater than diag of graph