Logistic Regression

Concept and Application in Data Science

Contents

Contents

Introduction to Logistic Regression

Introduction to Logistic Regression

- In linear regression, the Y variable is always a continuous variable. If suppose, the Y variable was categorical, you cannot use linear regression model on it.

So, what would you do when the Y is a categorical variable with 2 classes?

- Logistic regression can be used to model and solve such problems, also called as binary classification problems.
- A key point to note here is that Y can have 2 classes only and not more than that. If Y has more than 2 classes, it would become a multi class classification and you can no longer use the vanilla logistic regression for that.
- Yet, Logistic regression is a classic predictive modelling technique and still remains a popular choice for modelling binary categorical variables.
- Another advantage of logistic regression is that it computes a prediction probability score of an event. More on that when you actually start building the models.

Real World Examples Of Binary Classification Problems

Real World Examples Of Binary Classification Problems

Spam Detection : Predicting if an email is Spam or not

Credit Card Fraud : Predicting if a given credit card transaction is fraud or not

Health : Predicting if a given mass of tissue is benign or malignant

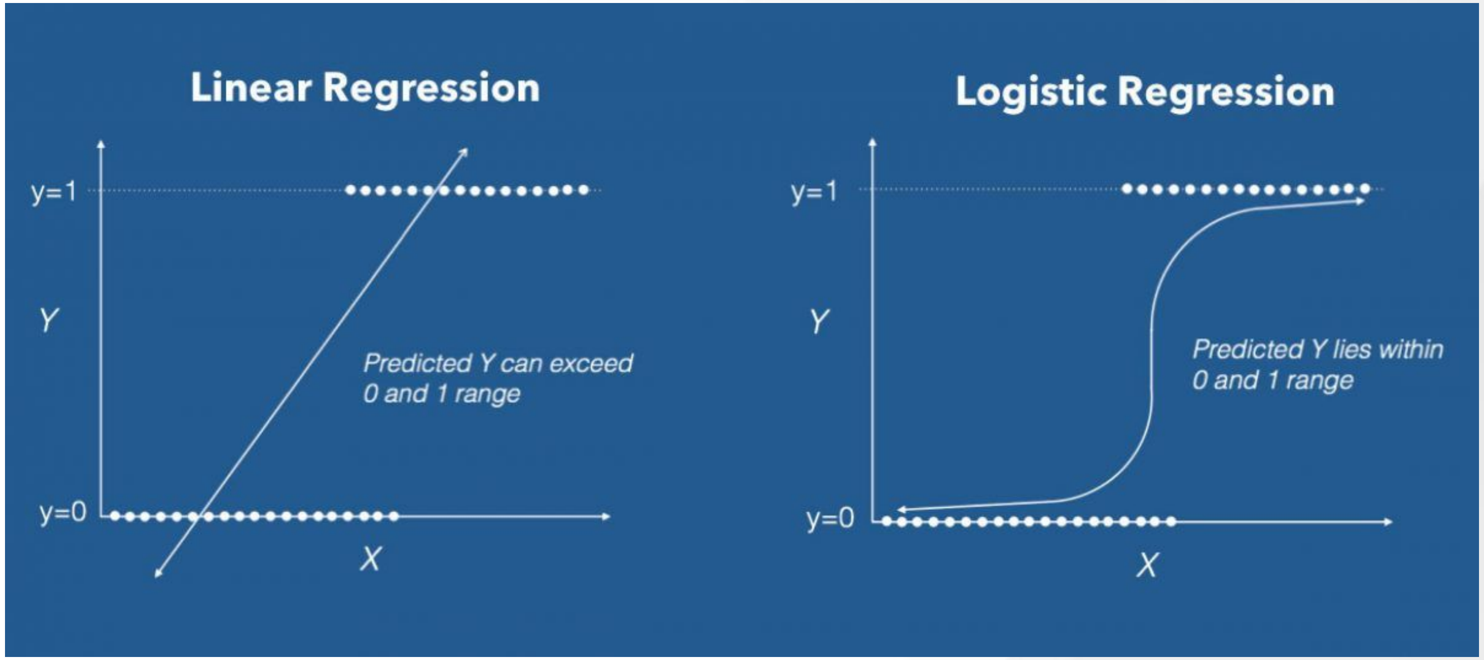
Marketing : Predicting if a given customer will respond to a campaign or not

Banking : Predicting if a customer will default on a loan.

Why Not Linear Regression?

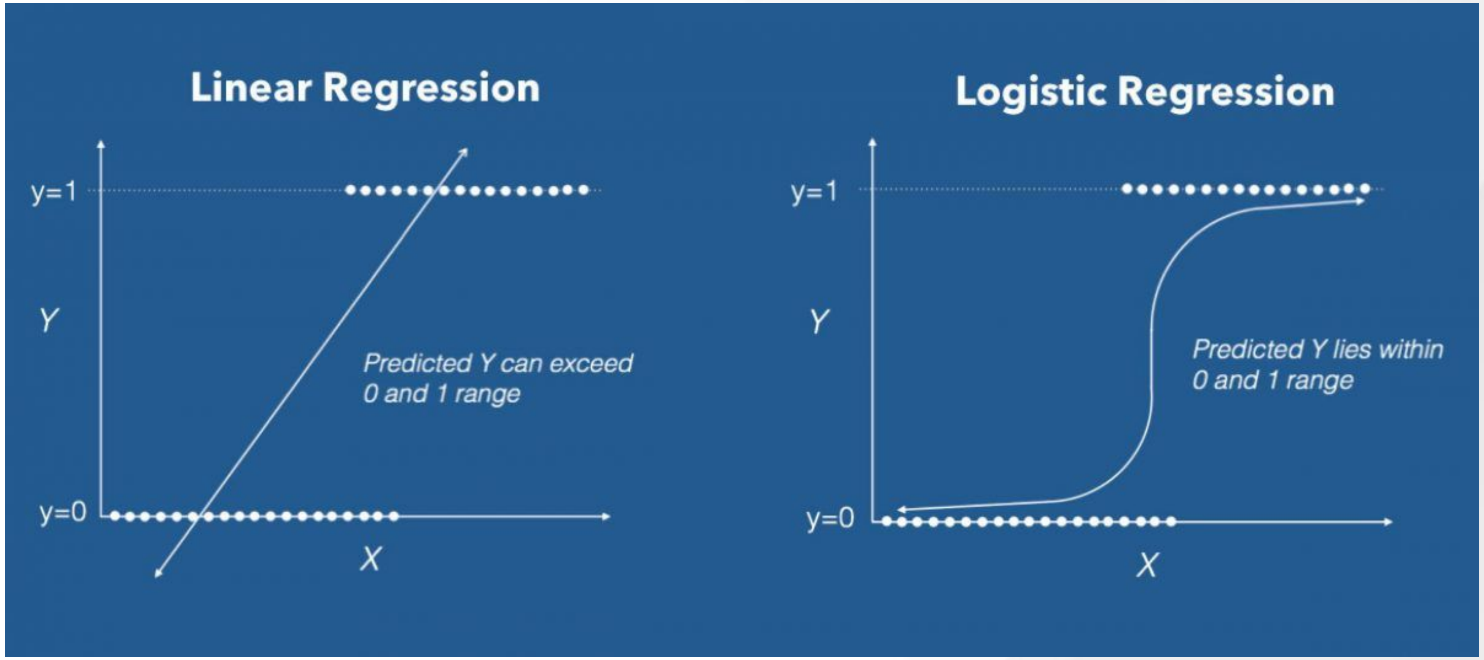
Why Not Linear Regression?

- When the response variable has only 2 possible values, it is desirable to have a model that predicts the value either as 0 or 1 or as a probability score that ranges between 0 and 1.
- Linear regression does not have this capability. Because, If you use linear regression to model a binary response variable, the resulting model may not restrict the predicted Y values within 0 and 1.



Why Not Linear Regression?

- This is where logistic regression comes into play.
- In logistic regression, you get a probability score that reflects the probability of the occurrence of the event.
- An event in this case is each row of the training dataset. It could be something like classifying if a given email is spam, or mass of cell is malignant, or a user will buy a product and so on.



The Logistic Regression Equation Derivation

The Logistic Regression Equation

- Logistic regression achieves this by taking the log odds of the event ln(P/1- P), where P is the probability of event. So, P always lies between 0 and 1.

$$\boxed{Zi = ln\left(\frac{Pi}{1 - Pi}\right) = \alpha + \beta 1x1 + ..+\beta nxn}$$

- Taking exponent on both sides of the equation gives:

$$\boxed{P_i = E(y = 1|x_i) = \frac{e^z}{1+ e^z} = \frac{e^{\alpha+ \beta_i x_i}}{1+ e^{\alpha+\beta_i x_i}}}$$

Derivation Of Logistic Regression Equation

- The fundamental equation of generalized linear model is:

$$\boxed{g(y) = \beta o + \beta(Age) \qquad \text{---- (a)}}$$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α,β,γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

Important Points

- GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
- The dependent variable need not to be normally distributed.
- It does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).
- Errors need to be independent but not normally distributed.

Derivation Of Logistic Regression Equation

- In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). As described above, g() is the link function. This function is established using two things: Probability of Success(p) and Probability of Failure(1-p). p should meet following criteria:
  - It must always be positive (since $p >= 0$)
  - It must always be less than equals to 1 (since $p <= 1$)

$$p = \exp(\beta o + \beta(Age)) = e\text{^}(\beta o + \beta(Age)) \quad \text{-------} \ (b)$$

$$p = \exp(\beta o + \beta(Age)) / \exp(\beta o + \beta(Age)) + 1 = e\text{^}(\beta o + \beta(Age)) / e\text{^}(\beta o + \beta(Age)) + 1 \quad \text{-----} \ (c)$$

$$p = e\text{^}y / 1 + e\text{^}y \quad \text{---} \ (d)$$

$$q = 1 - p = 1 - (e\text{^}y / 1 + e\text{^}y) \quad \text{---} \ (e)$$

Derivation Of Logistic Regression Equation

- On dividing, (d) / (e), we get:

$$\frac{p}{1 - p} = e^{y}$$

- After taking log on both side, we get:

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

- log(p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way.

$$\log\left(\frac{p}{1 - p}\right) = \beta_o + \beta(Age)$$

How To Deal With Class Imbalance?

How To Deal With Class Imbalance?

- Before building the logistic regressor, you need to randomly split the data into training and test samples.
- Since the response variable is a binary categorical variable, you need to make sure the training data has approximately equal proportion of classes.

Measuring Model Performance

How To Build Logistic Regression Model In R?

Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	benign
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10	9	7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	1	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign
1035283	1	1	1	1	1	1	3	1	1	benign
1036172	2	1	1	1	2	1	2	1	1	benign
1041801	5	3	3	3	2	3	4	4	1	malignant

- This is a breast cancer data set available in “mlbench” package.
- The goal here is to model and predict if a given specimen (row in dataset) is benign or malignant, based on 9 other cell features. So, let's load the data and keep only the complete cases.
- The dataset has 699 observations and 11 columns. The Class column is the response (dependent) variable, and it tells if a given tissue is malignant or benign.

Performance Of Logistic Regression Model

- **AIC (Akaike Information Criteria)** – The analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.
- **Null Deviance and Residual Deviance** – Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.
- **Confusion Matrix**: It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:

		Predicted	
		Good	Bad
Actual	Good	True Positive (d)	False Negative (c)
	Bad	False Positive (b)	True Negative (a)

Calculating model accuracy

$$\frac{\text{True Positive} + \text{True Negatives}}{\text{True Positive} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

ROC Curve

- **ROC Curve**: Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity).
- For plotting ROC, it is advisable to assume $p > 0.5$ since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of $p > 0.5$.
- The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve.
- Higher the area under curve, better the prediction power of the model.
- Here is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals 0. This curve will touch the top left corner of the graph.

ROC Curve

Note:

- For model performance, you can also consider likelihood function. It is called so, because it selects the coefficient values which maximizes the likelihood of explaining the observed data.
- It indicates goodness of fit as its value approaches one, and a poor fit of the data as its value approaches zero.

Thank You!