Assignment No.

Title: Logistic Regression

Problem Definition: To build a model in order to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset using Logistic Regression.

Objectives: Understand the implementation of the Logistic Regression model.

Outcomes: Predicting whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset using Logistic Regression.

Prerequisite: : Python 3, Jupyter Notebook, Operating System: Ubuntu / Windows.

Theory:

1.What is Logistic Regression?

Logistic Regression is a Classification algorithm. It is used to predict a binary outcome (1/0,Yes/No, True / False) given a set of independent variables. To represent binary/categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to logit function.

2.Derivation of Logistic Regression Equation

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (GLM).

The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

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.

Example:

We are provided a sample of 1000 customers. We need to predict the probability whether a customer will buy (y) a particular magazine or not. As you can see, we've a categorical outcome variable, we'll use logistic regression.

To start with logistic regression, first write the simple linear regression equation with dependent variable enclosed in a link function:

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

Note: Let 'Age' be an independent variable.

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:

1. It must always be positive (since p >= 0)
2. It must always be less than equals to 1 (since p <= 1)

Now, we'll simply satisfy these 2 conditions and get to the core of logistic regression. To establish link function, we'll denote g() with 'p' initially and eventually end up deriving this function.

Since probability must always be positive, we'll put the linear equation in exponential form. For any value of slope and dependent variable, exponent of this equation will never be negative.

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

To make the probability less than 1, we must divide p by a number greater than p. This can simply be done by:

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

Using (a), (b) and (c), we can redefine the probability as:

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

where p is the probability of success. *This (d) is the Logit Function*

If p is the probability of success, 1-p will be the probability of failure which can be written as:

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

where q is the probability of failure

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

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

After taking log on both side, we get,

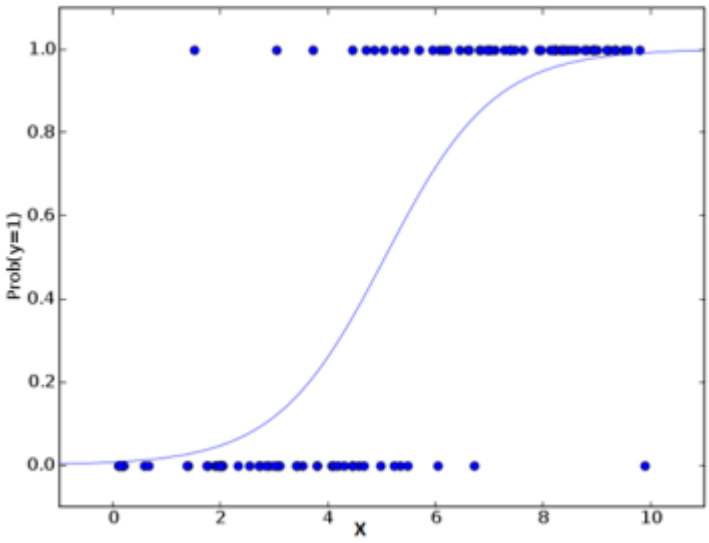
$$\log\left[\frac{p}{1 - p}\right] = y$$

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.

After substituting value of y, we'll get:

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

This is the equation used in Logistic Regression. Here (p/1-p) is the odd ratio. Whenever the log of odd ratio is found to be positive, the probability of success is always more than 50%. A typical logistic model plot is shown below. You can see probability never goes below 0 and above 1.



3.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.

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

You can calculate the **Accuracy** of your model with:

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

From confusion matrix, **Specificity** and **Sensitivity** can be derived as illustrated below:

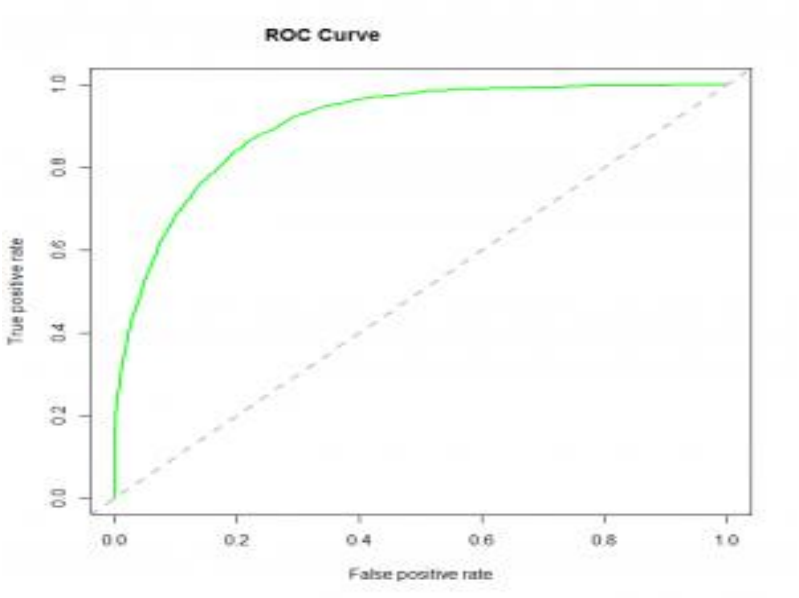
$$\left.\begin{aligned}\text{True Negative Rate (TNR), specificity} &= \frac{A}{A+B}\\ \text{False Positve Rate (FPR), } 1-\text{specificity} &= \frac{B}{A+B}\end{aligned}\right\}\text{sum to 1}$$

$$\left.\begin{aligned}\text{True Positive Rate (TPR), sensitivity} &= \frac{D}{C+D}\\ \text{False Negative Rate (FNR)} &= \frac{C}{C+D}\end{aligned}\right\}\text{sum to 1}$$

Specificity and Sensitivity plays a crucial role in deriving ROC curve.

3. Receiver Operating Characteristic(ROC)

ROC curve 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. Below 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.



5.How to evaluate the performance of the model?

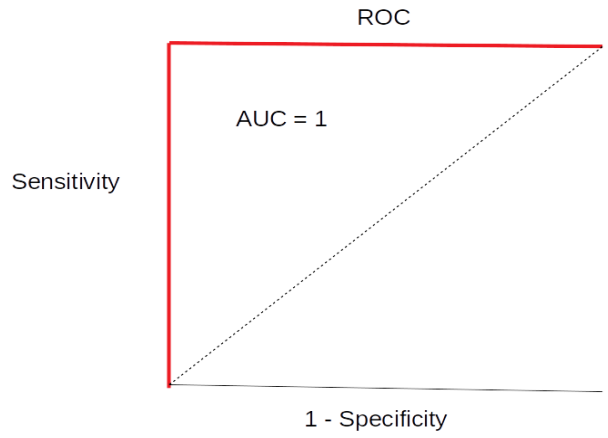
A model whose AUC = 1 is an ideal model, that means its accuracy is 100% and it can correctly predict all 0's as 0 and all the 1's as 1. And a model whose AUC = 0 is the worst model, that means its accuracy is 0% and its prediction is all wrong, it predicts 0's as 1 and the 1's as 0.

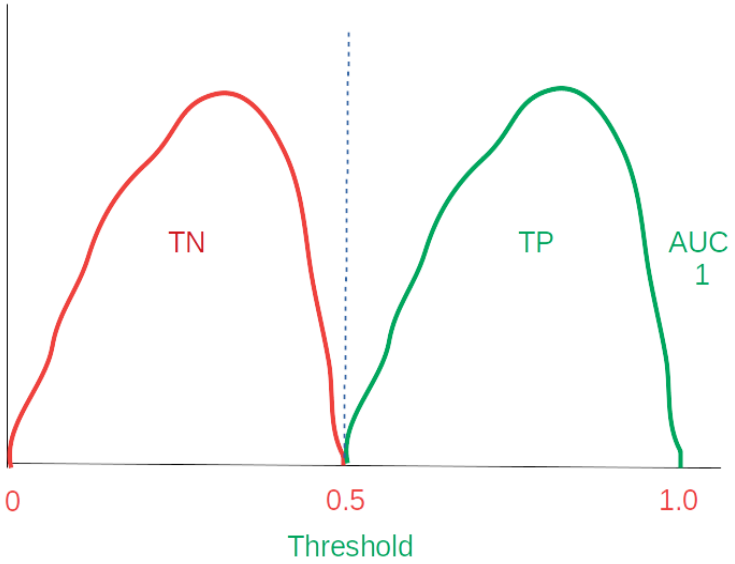
So a better model in which whose AUC is close to 1.

Let's plot the probability distribution graph.

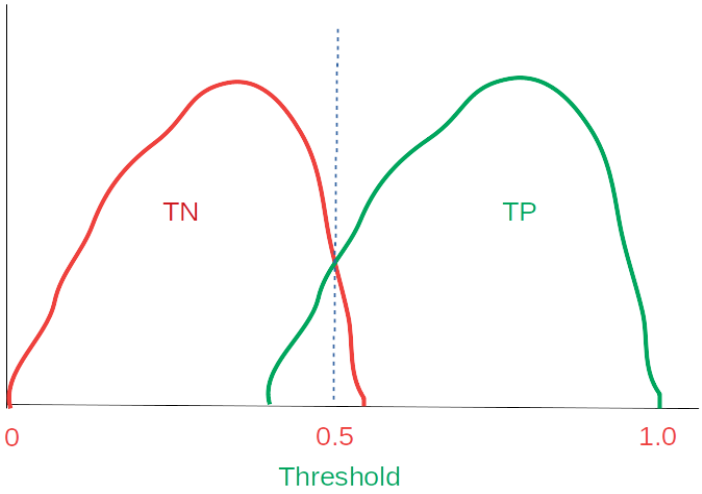
Note: Red Curve is for the negative class (Customer who won't buy a magzine.)

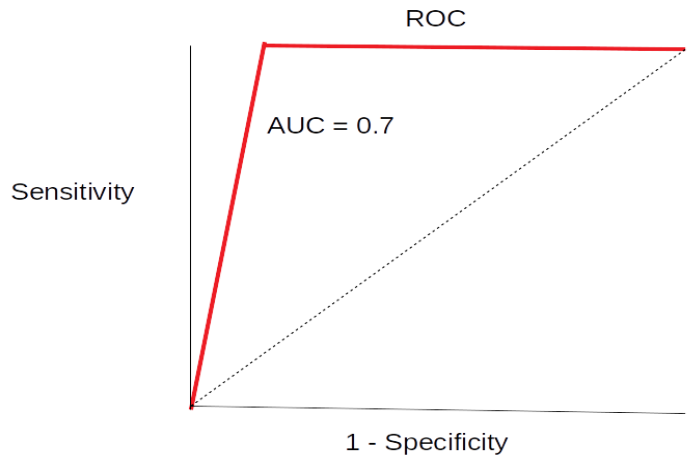
Green Curve is for the positive class (Customer who will buy a magzine.)



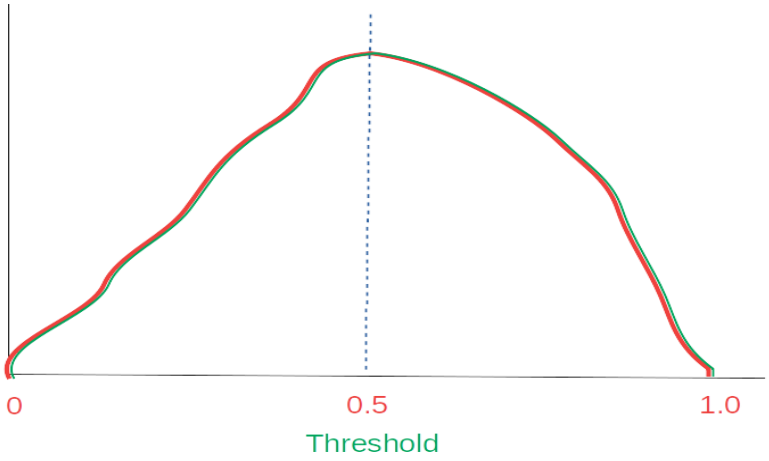


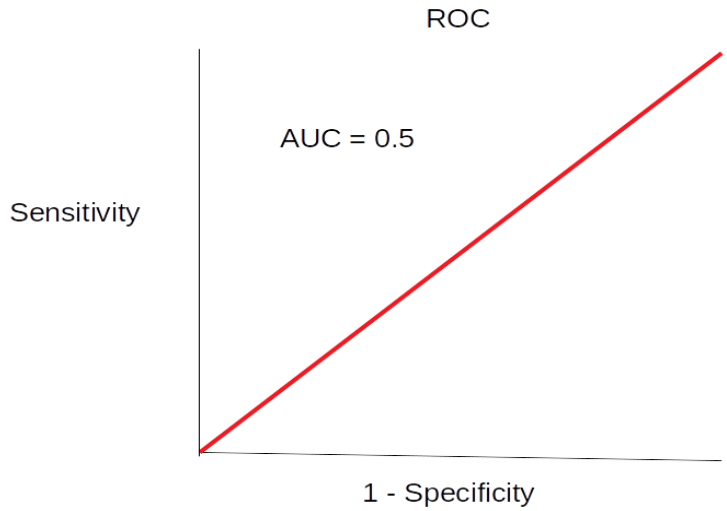
This is an ideal condition when the two curves are not overlaps, that means our model can perfectly distinguish the classes. And its AUC is 1.





When two curves overlap, it means some error is introduced, our model predicts some wrong 1's and 0's. But suppose if we have some other model whose AUC is greater than 0.7, than our second model is better than the first.





When two curves completely overlap, that means our model is not able to distinguish between the positive class and negative class. In this case, AUC is nearly equal to 0.5.

Conclusion : Hence, in this assignment of Logistic regression we have diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.