greatlearning

Logistic Regression

Logistic Regression - Topics

- Introduction to Logistic Regression
- Logit function in Logistic Regression
- Probability Examples
- Confusion Matrix
- F1 Score, gini index and ROC curve
- Pros and Cons of Logistic Regression
- Case study on Logistic Regression

Introduction to Logistic Regression

- In statistics, the logistic model is a statistical model that is usually taken to apply to a binary dependent variable.

- In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model.

- In Logistic Regression, the dependent variable is binary rather than continuous and it can also be applied to ordered categories (ordinal data).

The term “Odds”

- Popular in horse races, sports, gambling, epidemiology,
- Instead of talking about the *probability* of winning or contacting a disease, people talk about the *odds* of winning or contacting a disease
- How are these two different?

Additional content: Not part of the video but will be useful for getting an industry perspective

Logit function in Logistic Regression

In logistic regression, the dependent variable is a logit, which is the natural log of the odds, that is,

$$odds = \frac{P}{1-P}$$

$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

Additional content: Not part of the video but will be useful for getting an industry perspective

Odds vs Probability

greatlearning

-
- What is probability of A– P(A)?
- $Odds\ ratio = \frac{Probability\ of\ event\ occuring}{Probability\ of\ event\ not\ occuring}$
- Odds ratio $= \frac{P}{1-P}$
- Probability$= \frac{Odds\ ratio}{1+Odds\ ratio}$

Additional content: Not part of the video but will be useful for getting an industry perspective

Math behind Logistic Regression

So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find

logit(P) = a + bX,

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Math behind Logistic Regression

- Predict likelihood or probability
- Predicted value - >0 and <1
- Use of sigmoid function to achieve this

$$Probability\ (P) = \frac{e^z}{1 + e^z}$$

$z = \beta_0 + \beta_1 x$

$$Odds\ Ratio = \frac{P}{1 - P}$$

Substituting for P, $Odds\ Ratio = \frac{P}{1-P} = e^z = e^{(\beta_0 + \beta_1 x)}$

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

Log(Odds) takes the form of linear regression
intercept β_0 and slope β_1
Generalized Linear Model
β_0 and slope β_1 estimated using maximum likelihood estimation

Equation of logistic regression

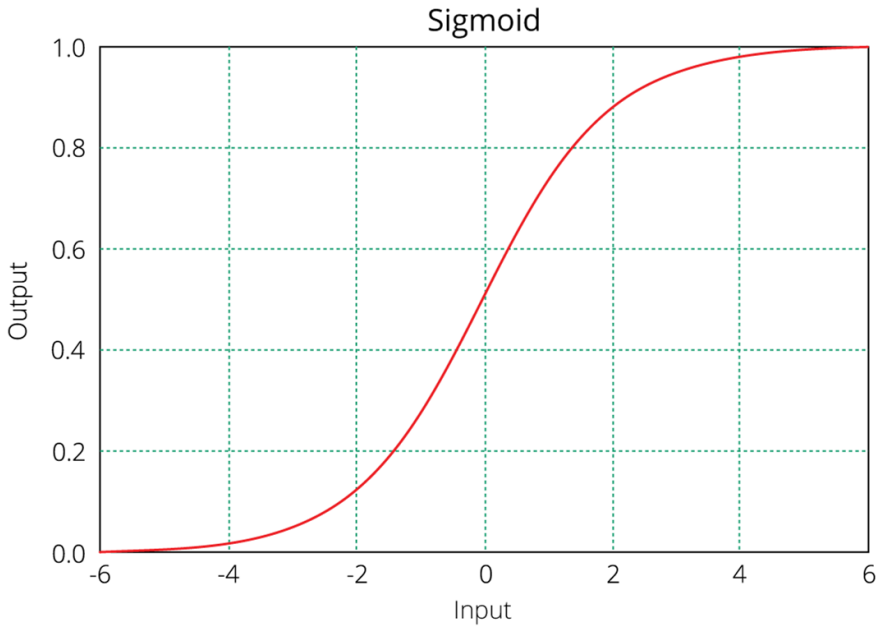
greatlearning

log - odds or **odds ratio** or **logit** function and is the link function for Logistic Regression

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Regression intercept & coefficient

This link function follows a sigmoid function which limits its range of probabilities between 0 and 1.

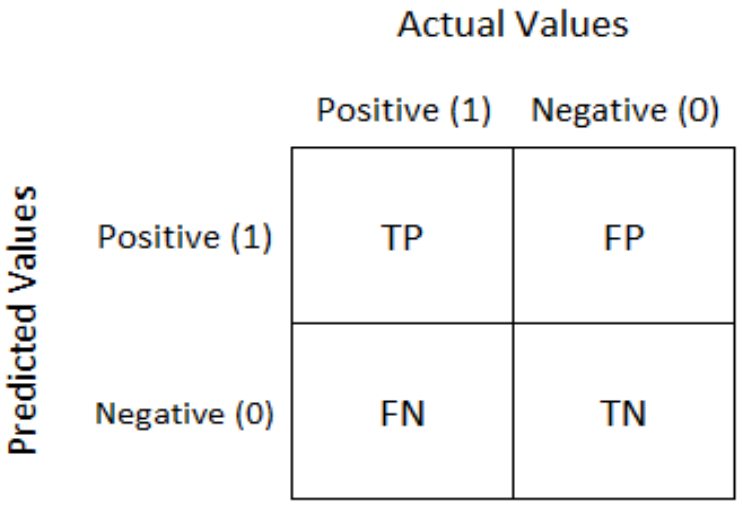


Probability Examples

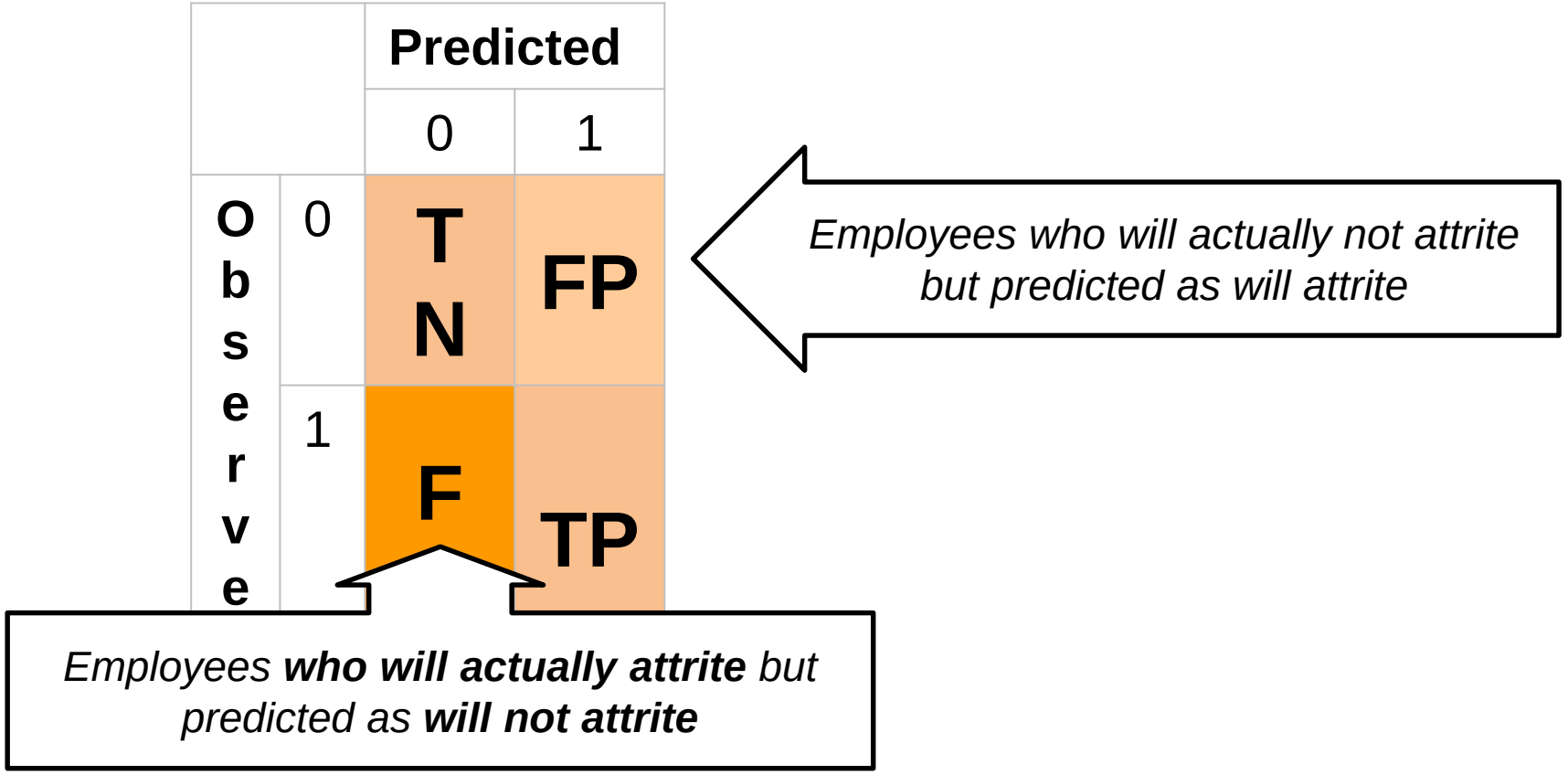
Given the value of predictor (variable x), the model estimates the probability that the new data point belongs to a given class “A”. Probability values can range between 0 and 1.

Confusion Matrix

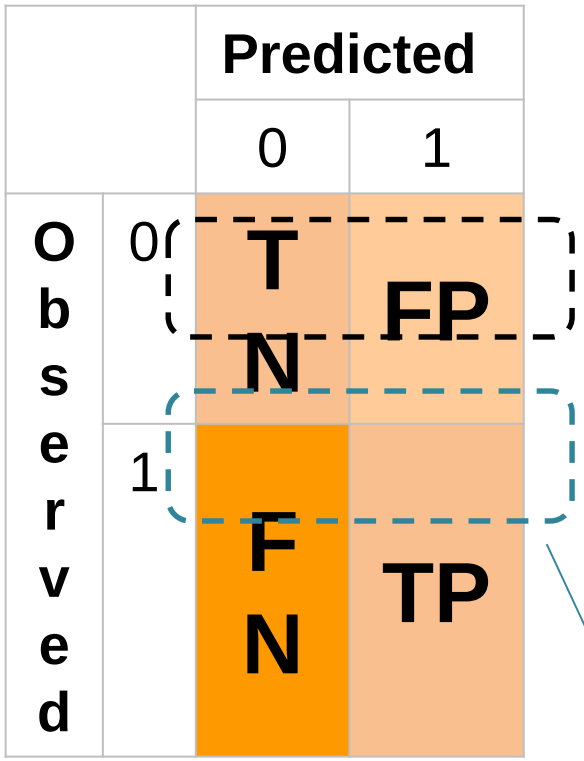
Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.



Confusion matrix



Confusion Matrix



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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall\ or\ Sensitivity = \frac{TP}{TP + FN}$$

greatlearning

Why accuracy is not a good model performance measure?

Total customers – 100
of customers who churn - 8

Churn rate – 8%

Model predicts that no one churns

Accuracy – 92%

Misses out critical customers who churn

		Predicted	
		0	1
Observed	0	TN = 92	FP
	1	FN = 8	TP

F1 Score

A single metric is not sufficient for the evaluation of classification models. We have seen that we need to use recall and precision together along with accuracy to evaluate our model.

Let us consider another metric that puts together the recall and precision metrics. We call it F1 Score.

F1 Score = 2(precision*recall)/precision + recall

-which is the harmonic mean of the two metrics.

The F1 score can also be used to evaluate the model.

ROC and Gini Coefficient and Threshold

greatlearning

- Roc is a curve which allows us to compare models.
- It is plot between TPR(true positive rates) and FPR(false positive ratio).
- The area under the ROC curve (AUC) is a measure of the how good a model is.

Gini Coefficient:

- It is also used to measure the goodness of a fit.
- It is the ratio of areas in a roc curve and is scaled version of the AUC.
- GI = 2*AUC -1

Pros and Cons of logistic regression

Pros:

- It is a model that gives probabilities.
- It can be easily scaled to multiple classes.
- It is very quick to train and very fast at classifying unknown records.

Cons:

- The classifies constructs liner boundaries.
- Assumes that the variables are independent.
- Interpretation of coefficients is difficult.

Hands on exercise on Logistic Regression

greatlearning

HappyLearning

Questions?