W7 Formulae

Thursday, 17 November 2022 22:11

Logistic Regression

Logistic regression is used to predict a dependent categorical variable.

$$\boldsymbol{Odds\left(success\right) = \frac{P(Y=1)}{P(Y=0)}}$$

In logistic regression, we use log of odds.

$$\log(Odds) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

then

$$Odds = e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}$$

$$P(Y=1) = \frac{Odds}{1 + Odds} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4}}$$

- Above expression will give us probabilities. But, we want to make predictions in $0-1$ (yes/no) format.
- So, we set a threshold (or cut-off) which is a number.

If $P(Y=1) >$ cut-off:
  $\hat{Y} = 1$

Else:
  $\hat{Y} = 0$

- In logistic regression, if an explanatory variable increases by 1 unit, then the odds of $Y=1$ increases by a factor of e^{β}.

Evaluation of logistic regression model

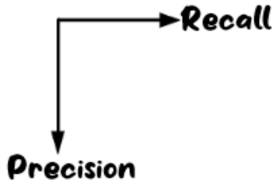
Confusion Matrix

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{Total number of observations}}$$

For predicting 1	For predicting 0
$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$	$\text{Precision} = \frac{\text{TN}}{\text{TN}+\text{FN}}$
$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$	$\text{Recall} = \frac{\text{TN}}{\text{TN}+\text{FP}}$

Tip to Remember:

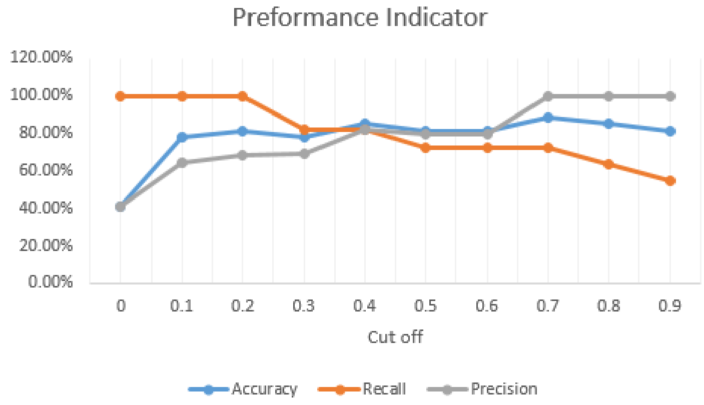


- The parameters are **larger the better** type.

How to find optimal cut-off?

At different cut off-values, we calculate these three performance indicators: accuracy, precision, recall, and then compare the results.

Example:



In the above analysis, optimal cut-off $= 0.4$