Logistic Regression

Agenda

After this session, you will know:

- Classification with Linear Regression
- Probability and Odds
- Confusion Matrix
- Intro to Logistic Regression
- Interpreting Coefficients in Logistic Regression
- Assumptions of Logistic Regression

Business Problem | Credit Risk Detection

Bank customers seek loans from the bank promising to repay the loan in installments over a determined period of time and with some interest on the amount.

However, banks are always at a risk because many customers might not be able to pay their loans back. This can cause big losses to the bank.

Therefore, predicting credit risk is of utmost importance for the bank where the bank analyzes customers’ information and credit history before deciding to grant a loan.

Logistic Regression can be used to build a predictive modeling to predict how likely a customer is to default the repayment of the loan.

Revisiting Regression

Regression analysis is used to:

Predict the value of a dependent variable based on the value of at least one or more independent variable (s)

Explain the impact of changes in an independent variable on the dependent variable

Dependent Variable

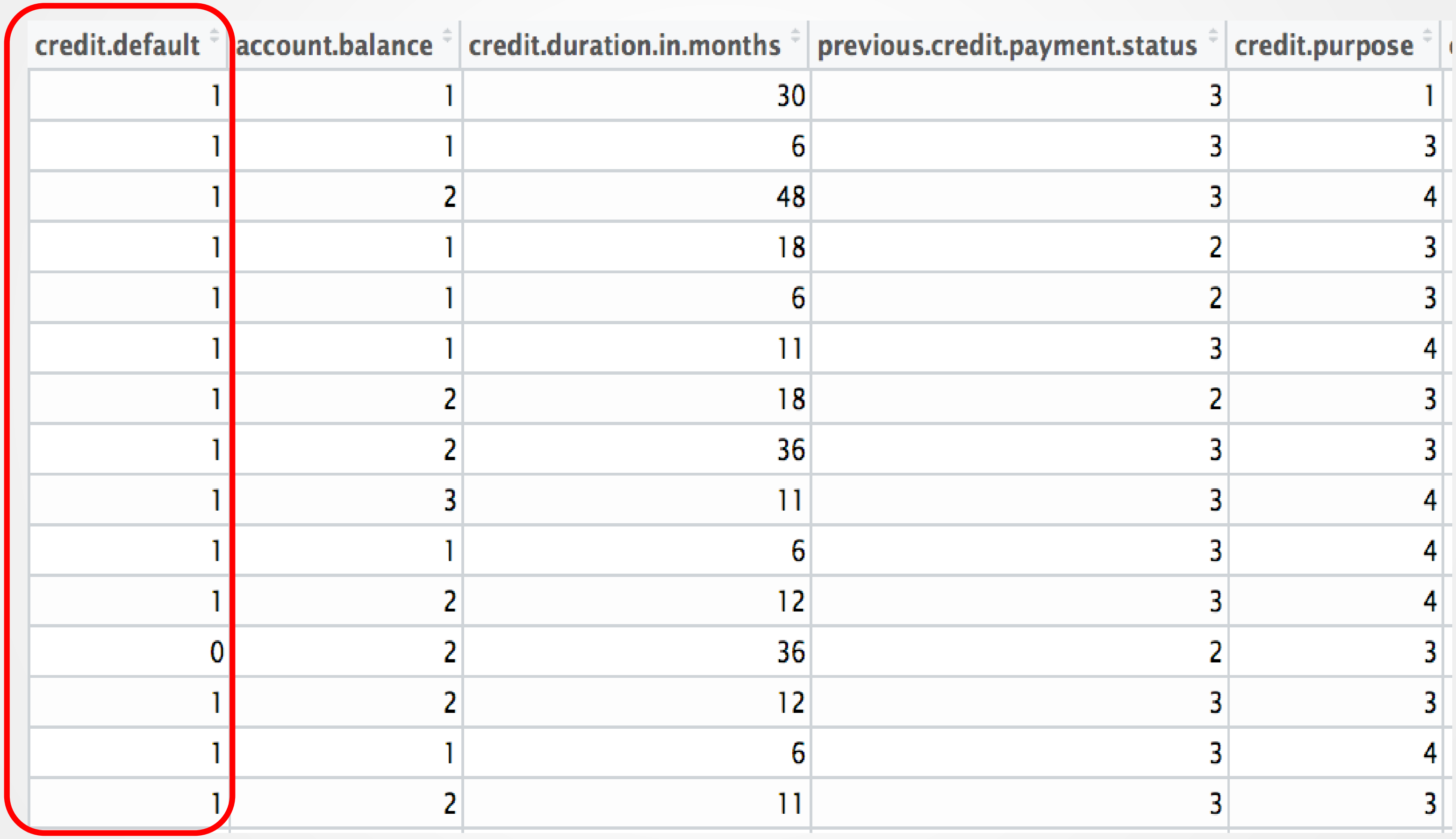
The variable we wish to explain usually denoted by Y

Independent Variable(s)

The variable(s) used to explain the dependent variable and denoted by X

Classifying with Linear Regression

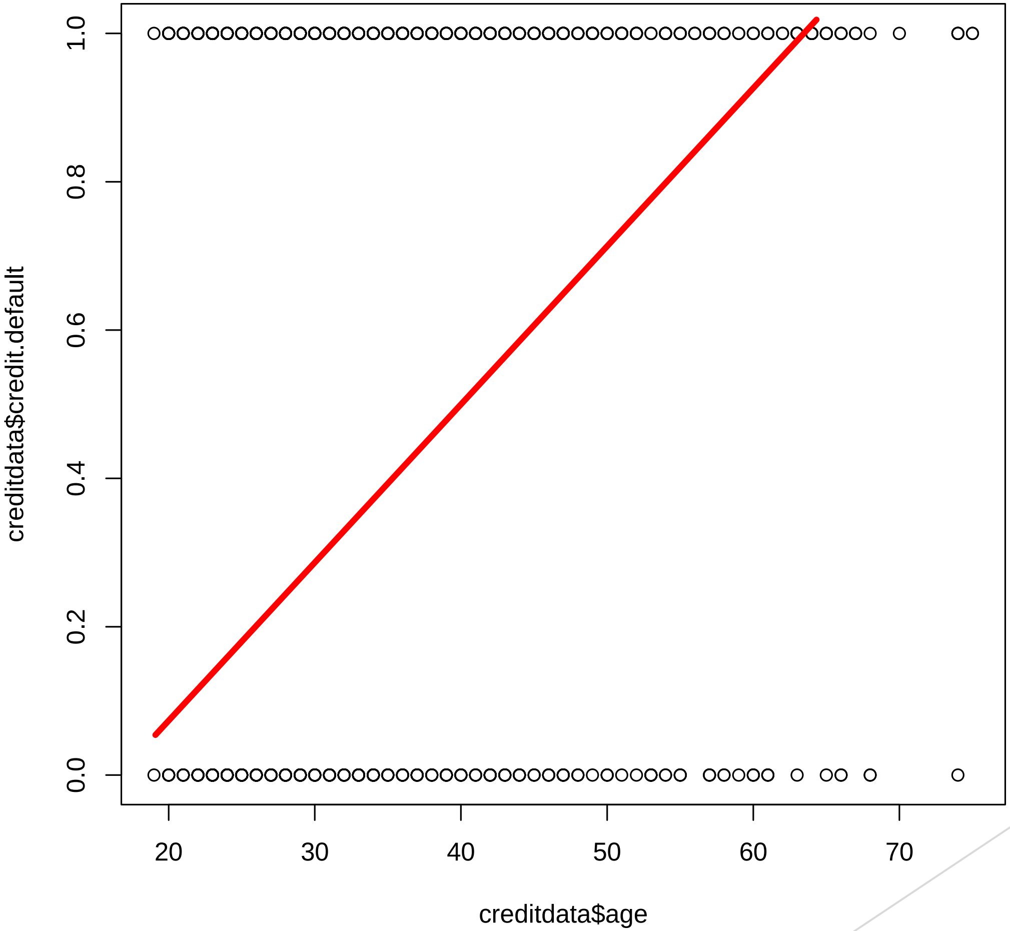
Our output variable is Credit Default in the dataset which is either 0 or 1



credit.default	account.balance	credit.duration.in.months	previous.credit.payment.status	credit.purpose
1	1	30	3	1
1	1	6	3	3
1	2	48	3	4
1	1	18	2	3
1	1	6	2	3
1	1	11	3	4
1	2	18	2	3
1	2	36	3	3
1	3	11	3	4
1	1	6	3	4
1	2	12	3	4
0	2	36	2	3
1	2	12	3	3
1	1	6	3	4
1	2	11	3	3

Classifying with Linear Regression

Trying to fit in a best fit regression line…..



Defaulted

Not Defaulted

Where is the Problem?

Dependent variable is limited to the [0 & 1] because we have 2 classes: default or no-default.

Linear Regression is designed to solve the problem of minimizing the Squared Error, which does not seem to be an appropriate fit in this case.

Probability & Odds

$$\text{Probability} = \frac{\text{Outcome of Interest}}{\text{All possible outcomes}}$$

$$\text{Odd} = \frac{\text{P(Occurring)}}{\text{P(Not Occurring)}}$$

Fair coin flip:
P(heads) = 1/2
= 0.5

Fair die roll:
P(2 or 4) = 2/6
= 0.33

Playing cards:
P(Heart) = 13/52
= 0.25

Fair coin flip:
odds(heads) = 0.5/0.5
= 1

Fair die roll:
odds(2 or 4) = .33/.66
= 0.5

Playing cards:
odds(Heart) = 0.25/0.75
= 0.33

Odds

Recall the definitions of **Odds**:

$$\boxed{odds = \frac{p}{1-p}}$$

The odds has a range of 0 to ∞ with values greater than 1 associated with an event being more likely to occur than to not occur and values less than 1 associated with an event that is less likely to occur than not occur

Odds (contd.)

The **logit** is defined as the log of the odds:

$$\boxed{\ln(odds) = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p)}$$

This transformation creates a variable with a range from -∞ to +∞

- It solves the problem we encounter in fitting a linear model to probabilities
- As probabilities (the dependent variable) only range from 0 to 1, we can get linear predictions that are outside of this range

Odds Ratio

Any odds ratio, by definition, is a ratio of two odds, written here as Odds_Y divided by Odds_X, in which the subscripts indicate two individuals or two groups of individuals being compared.

$$\boxed{\text{Odds Ratio(OR)} = \frac{\text{Odds}_Y}{\text{Odds}_X}}$$

So if the outcome is the same in both groups the ratio will be 1, which implies there is no difference between the two group of individuals.

However:

If the OR is > 1 then group X is better than group Y

If the OR is < 1 then group Y is better than group X

Why Use Odds Ratio?

The problem is that Probability and Odds have different properties that give Odds some advantages

- For example, in logistic regression the Odds Ratio represents the **constant effect** of a predictor X, on the likelihood that one outcome will occur
- In regression models, we often want a measure of the unique effect of each X on Y. If we try to express the effect of X on the likelihood of a categorical Y having a specific value through Probability, the effect is **not constant**
- That means there is no way to express in one number how X affects Y in terms of Probability. The effect of X on the probability of Y has different values depending on the value of X
- We will not be able to describe that effect in a single number using Probability (although it is intuitive) and will have to use Odds Ratio

Intro to Logistic Regression

Logistic regression is used to analyze relationships between a dichotomous dependent variable and categorical or numerical independent variables

1

Logistic regression combines the independent variables to estimate the probability that a particular event will occur

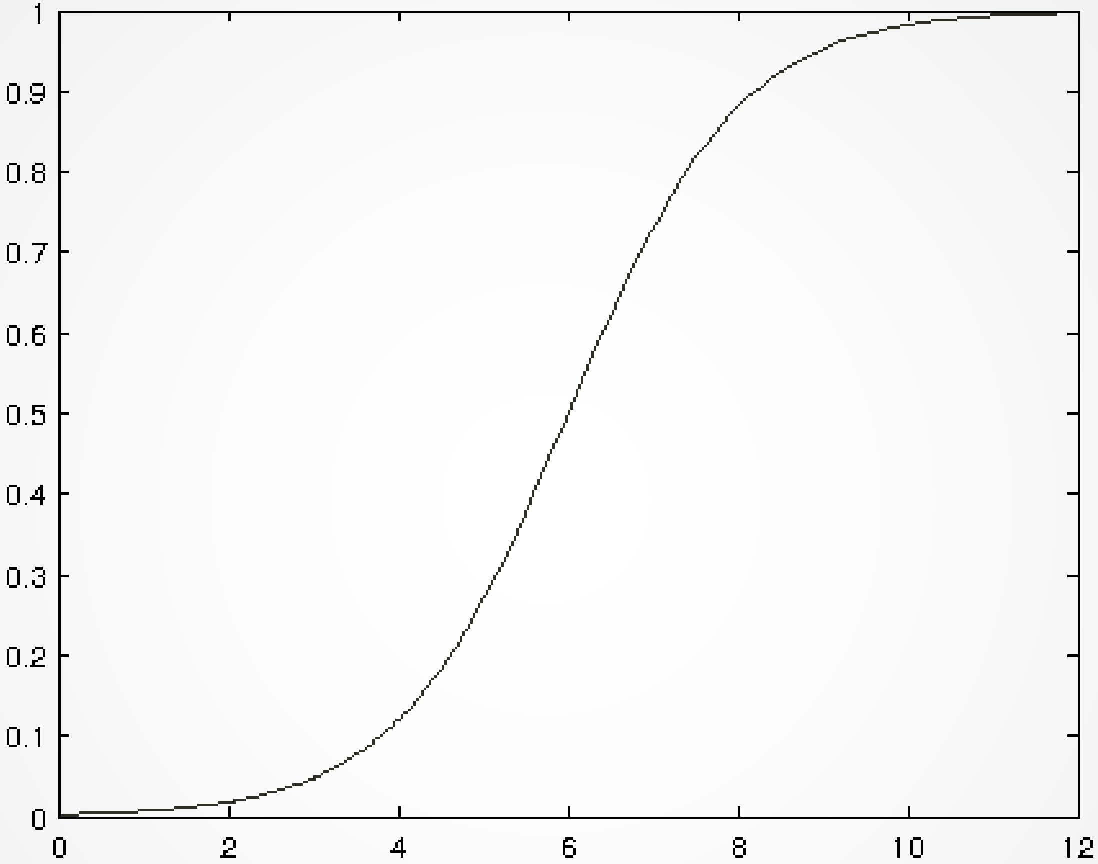
2

The value produced by logistic regression is a probability value between 0.0 and 1.0

If the probability for group membership in the modeled category is above some cut point (the default is 0.50), the subject is predicted to be a member of the modeled group

If the probability is below the cut point, the subject is predicted to be a member of the other group

Intro to Logistic Regression



$$f(x) = p(y=1) = e^x / (e^x + 1)$$

Logit Model

The "logit" model:

$$\boxed{\ln[p/(1-p)] = \beta_0 + \beta_1 X}$$

- p is the probability that the event Y occurs, p(Y=1)

  [range=0 to 1]

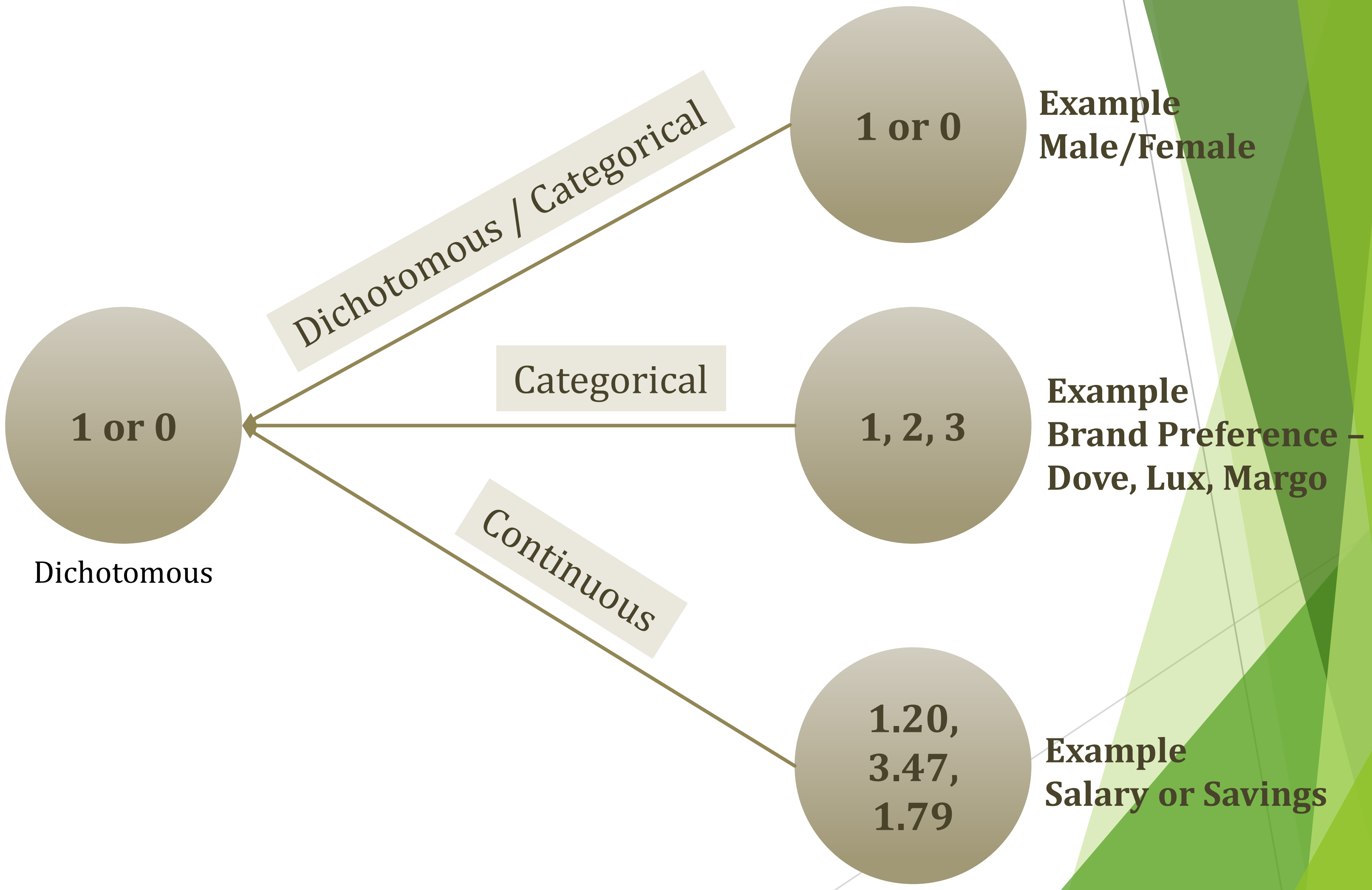
- p/(1-p) is the "odds ratio"

  [range=0 to ∞]

- ln[p/(1-p)]: log odds ratio, or "logit“

  [range=-∞ to +∞]

Generalized Linear Models



Response Variable

Predictor Variables

1 or 0

Dichotomous

Dichotomous / Categorical

Categorical

Continuous

1 or 0

Example
Male/Female

1, 2, 3

Example
Brand Preference –
Dove, Lux, Margo

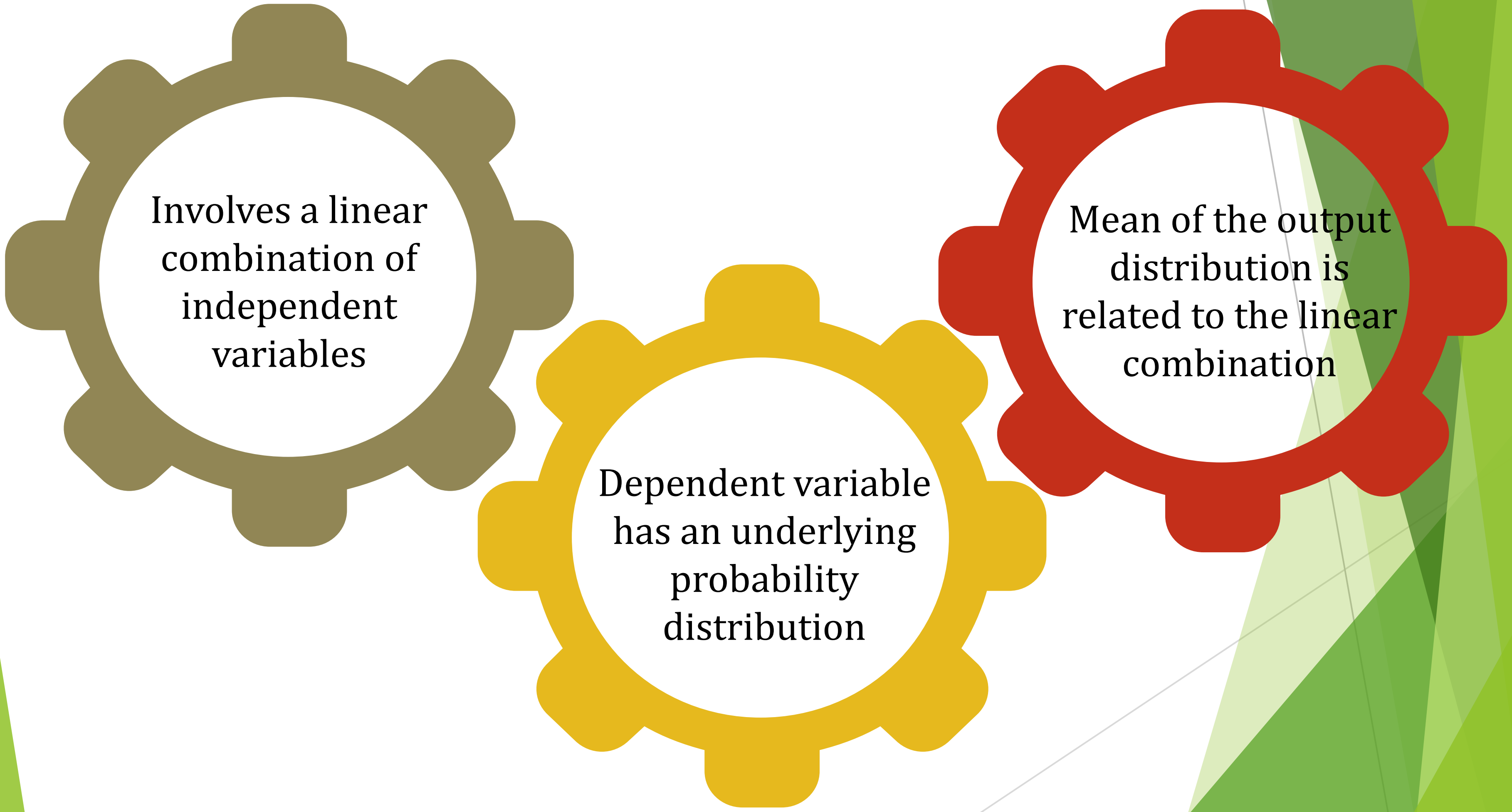
1.20,
3.47,
1.79

Example
Salary or Savings

Generalized Linear Models (GLM)

Logistic Regression belongs to a class of models known as GLM

Characteristics of GLM

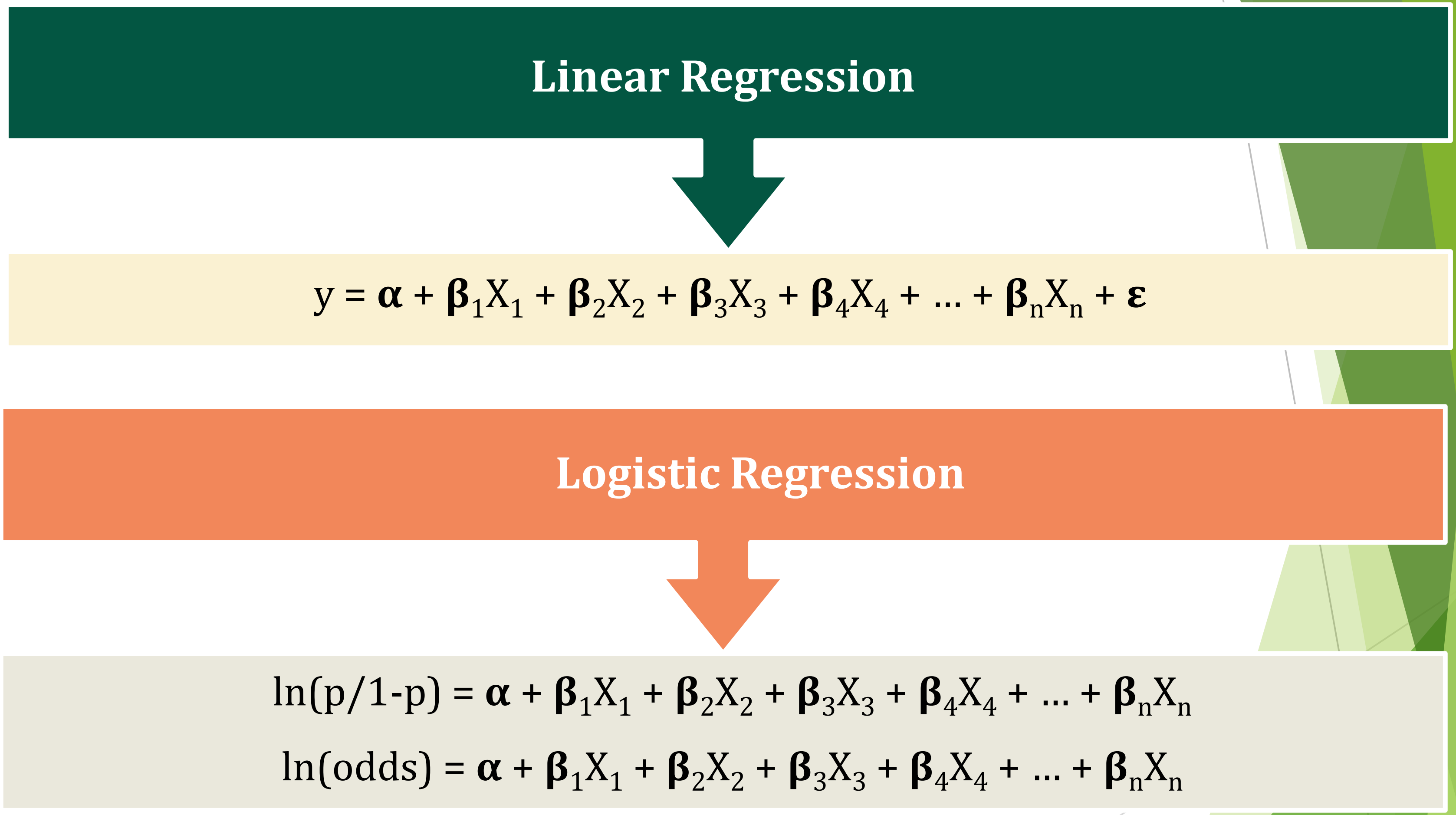


Involves a linear combination of independent variables

Dependent variable has an underlying probability distribution

Mean of the output distribution is related to the linear combination

The Logistic Regression Model



Linear Regression

$$y = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2 X_2 + \boldsymbol{\beta}_3 X_3 + \boldsymbol{\beta}_4 X_4 + \ldots + \boldsymbol{\beta}_n X_n + \boldsymbol{\varepsilon}$$

Logistic Regression

$$\ln(p/1\text{-}p) = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2 X_2 + \boldsymbol{\beta}_3 X_3 + \boldsymbol{\beta}_4 X_4 + \ldots + \boldsymbol{\beta}_n X_n$$

$$\ln(\text{odds}) = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2 X_2 + \boldsymbol{\beta}_3 X_3 + \boldsymbol{\beta}_4 X_4 + \ldots + \boldsymbol{\beta}_n X_n$$

Interpreting Coefficients in Logistic regression

01 RHS of both Linear Regression & Logistic Regression looks same except the error in case of Linear regression

02 In case of Logistic Regression, the LHS has the logit function

03 In Logistic regression, a unit increase in feature X_1, results in multiplying the odds ratio by an amount of e^x

04 If a coefficient is positive, then multiply the odds ratio by a number > 1

05 If a coefficient is negative, increasing the feature will shift the balance towards predicting class 0

Assumptions of Logistic Regression

Makes fewer assumptions than Linear Regression

Non-linear transformation of the logistic function

No assumption on normal distribution for residuals

No homoscedasticity assumption

Maximum Likelihood Estimation

In Linear Regression, we got our coefficients by *minimizing* the sum of squared error terms

For Logistic Regression, we do this by *maximizing* the likelihood of the data

The likelihood function is the product of all the individual likelihoods for each data point given by,

$$l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, .., \boldsymbol{\beta}_n) = \prod p(y=1) \,.\, \prod 1\text{-}p\ (y=1)$$

The idea is to choose our regression coefficients so that the likelihood function is maximized

Logistic Regression Variable Requirements

1. Logistic regression analysis requires that the dependent variable be dichotomous

2. Logistic regression analysis requires that the independent variables be numerical or dichotomous

3. If an independent variable is nominal level and not dichotomous, we need to dummy code the variable

4. Logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables

Confusion Matrix

	Positive Predicted	Negative Predicted
Positive Actual	True Positive	False Negative
Negative Actual	False Positive	True Negative

	Fire Predicted	No Fire Predicted
Fire Actual	True Positive	False Negative
No Fire Actual	False Positive	True Negative

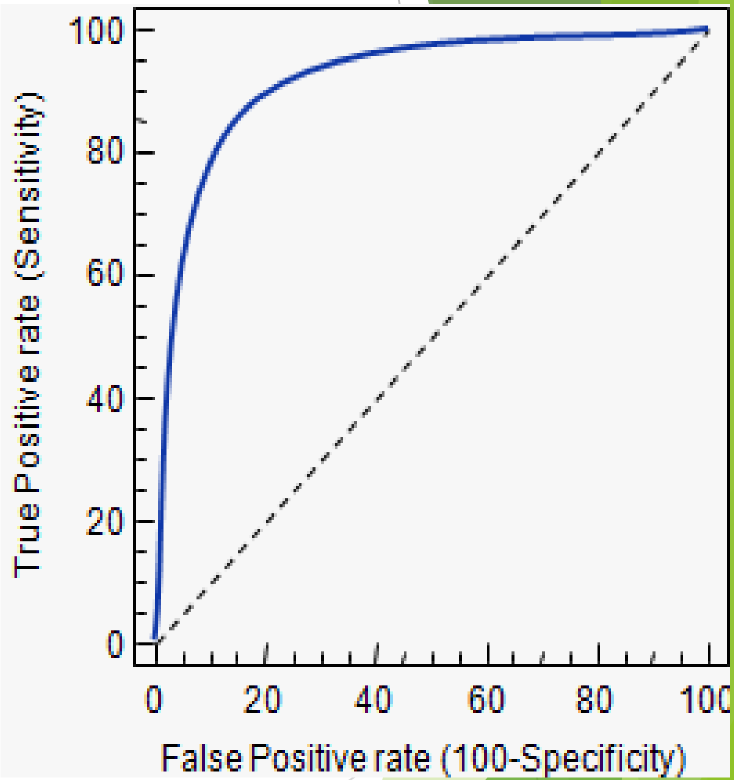
Receiver Operating Characteristic (ROC)

1 Tradeoff between sensitivity and specificity

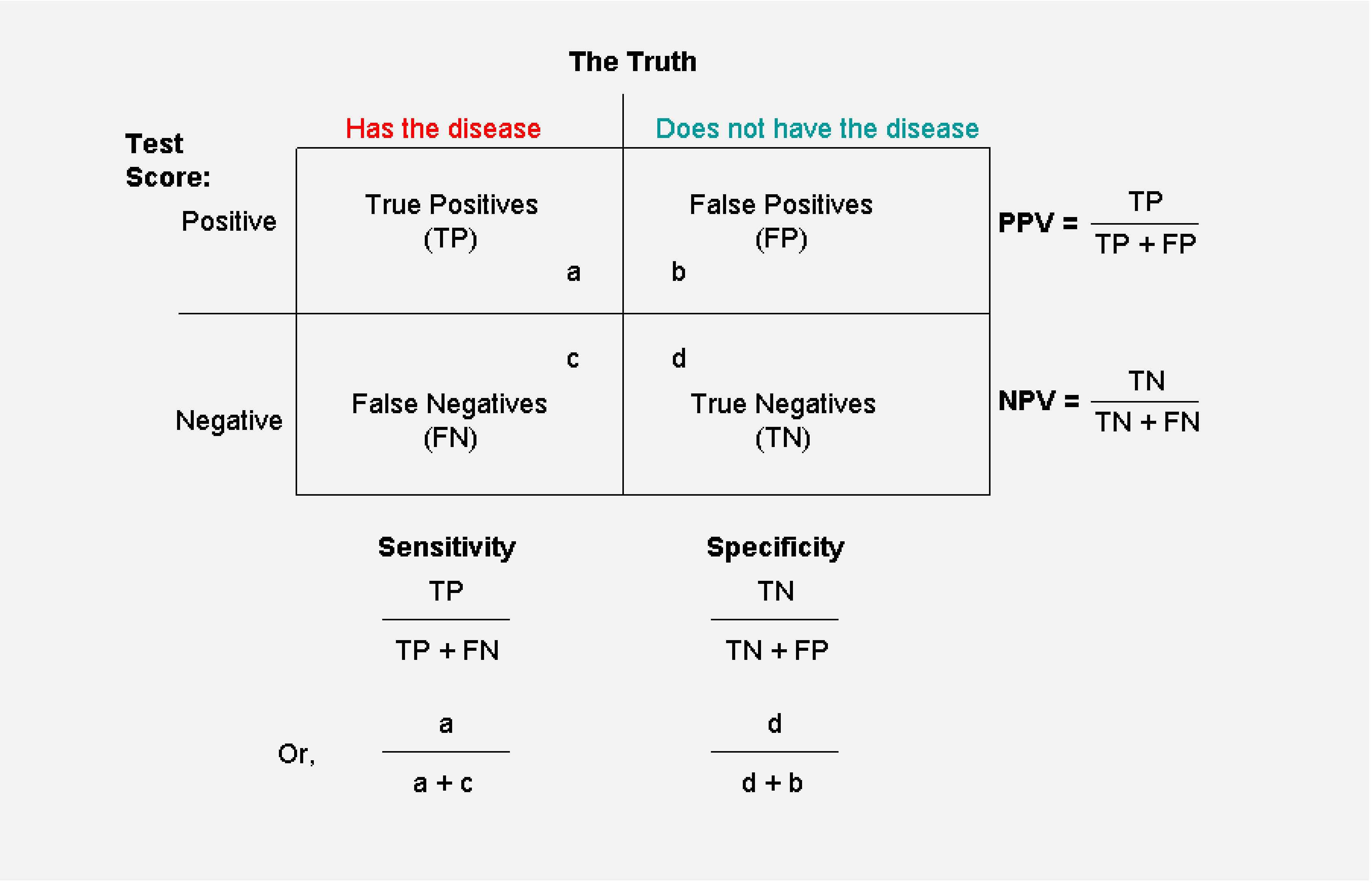
2 The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

3 The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test

4 The area under the curve is a measure of text accuracy



Sensitivity & Specificity



The Truth

Test Score:	Has the disease	Does not have the disease	
Positive	True Positives (TP) a	False Positives (FP) b	**PPV =** $\frac{TP}{TP + FP}$
Negative	c False Negatives (FN)	d True Negatives (TN)	**NPV =** $\frac{TN}{TN + FN}$

Sensitivity

$$\frac{TP}{TP + FN}$$

Specificity

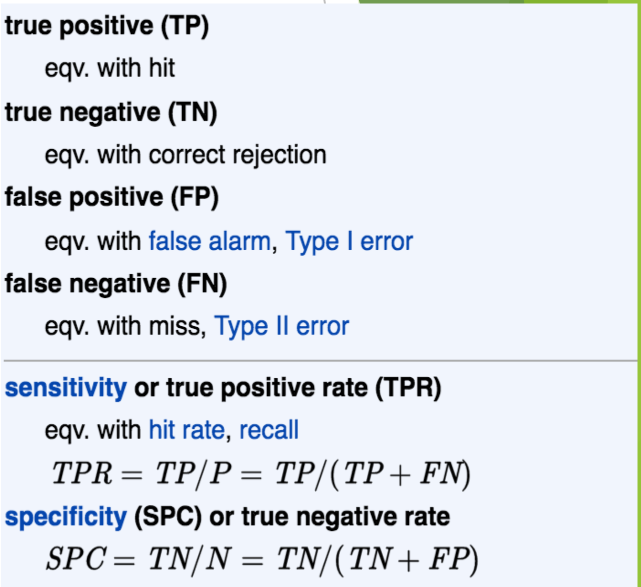
$$\frac{TN}{TN + FP}$$

Or, $\frac{a}{a + c}$ $\frac{d}{d + b}$

Sensitivity & Specificity

Imagine a study evaluating a new test that screens people for a disease. Each person taking the test either has or does not have the disease. The test outcome can be positive (classifying the person as having the disease) or negative (classifying the person as not having the disease). The test results for each subject may or may not match the subject's actual status. In that setting:

- True positive: Sick people correctly identified as sick
- False positive: Healthy people incorrectly identified as sick
- True negative: Healthy people correctly identified as healthy
- False negative: Sick people incorrectly identified as healthy



true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, Type I error

false negative (FN)
eqv. with miss, Type II error

sensitivity or true positive rate (TPR)
eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or true negative rate

$$SPC = TN/N = TN/(TN + FP)$$

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

In general, Positive = identified &negative = rejected.
Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected