Logistic Regression

Rina BUOY

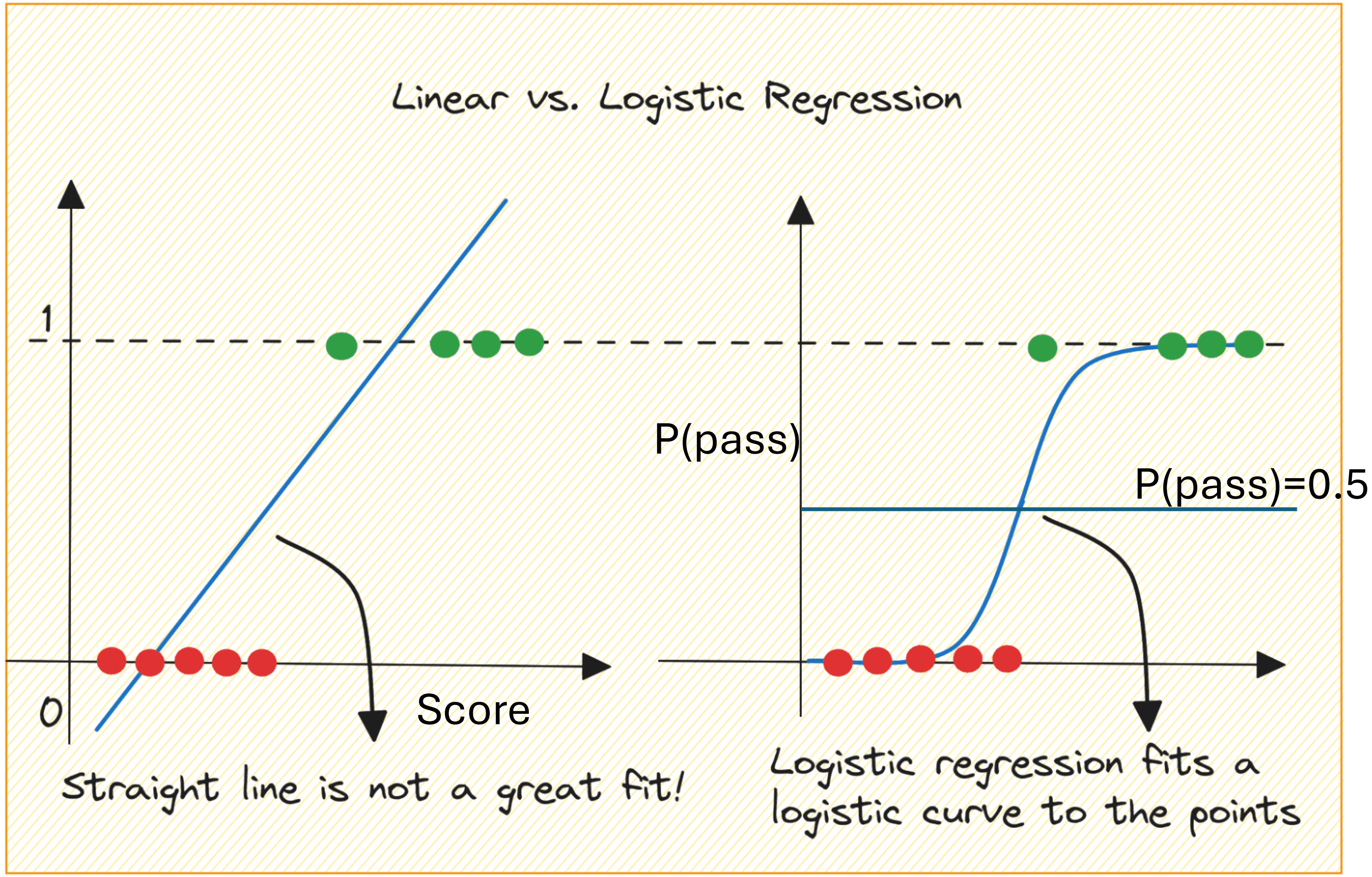
AMERICAN UNIVERSITY OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

Linear vs Logistic Regression

Score	Label
10	Fail (0)
60	Pass (1)
80	Pass (1)
...	...
100	Pass (1)

Model a linear relationship between score and outcome (pass/fail).

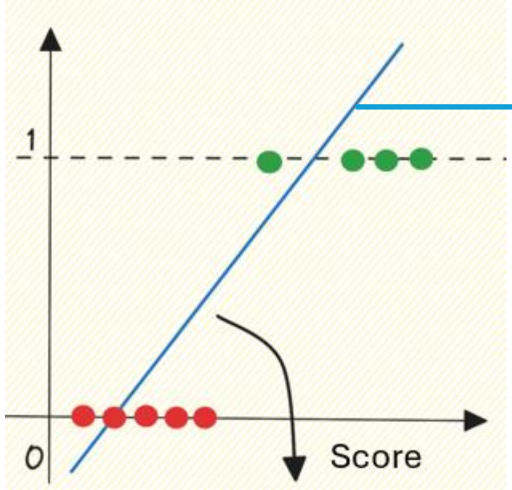


Sigmoid Function

- Use a sigmoid function to convert a real number (-inf,+inf) to a probability range (0,1).

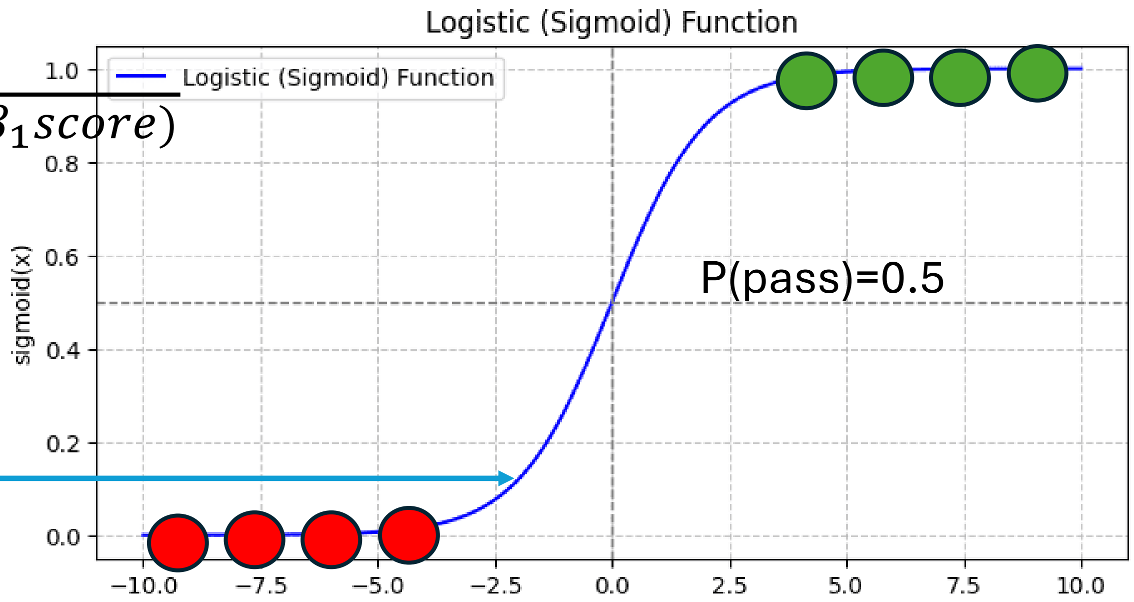
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\beta_0 + \beta_1 score$$



$$\frac{1}{1+e^{-(\beta_0+\beta_1 score)}}$$

P(pass)



$$\beta_0 + \beta_1 score$$

Connecting Probability, Odds, and Log Odds

- If the probability of passing the course is **70% (i.e., P(pass) = 0.5)**, **then P(fail) = 1.0 – P(pass) = 0.3**

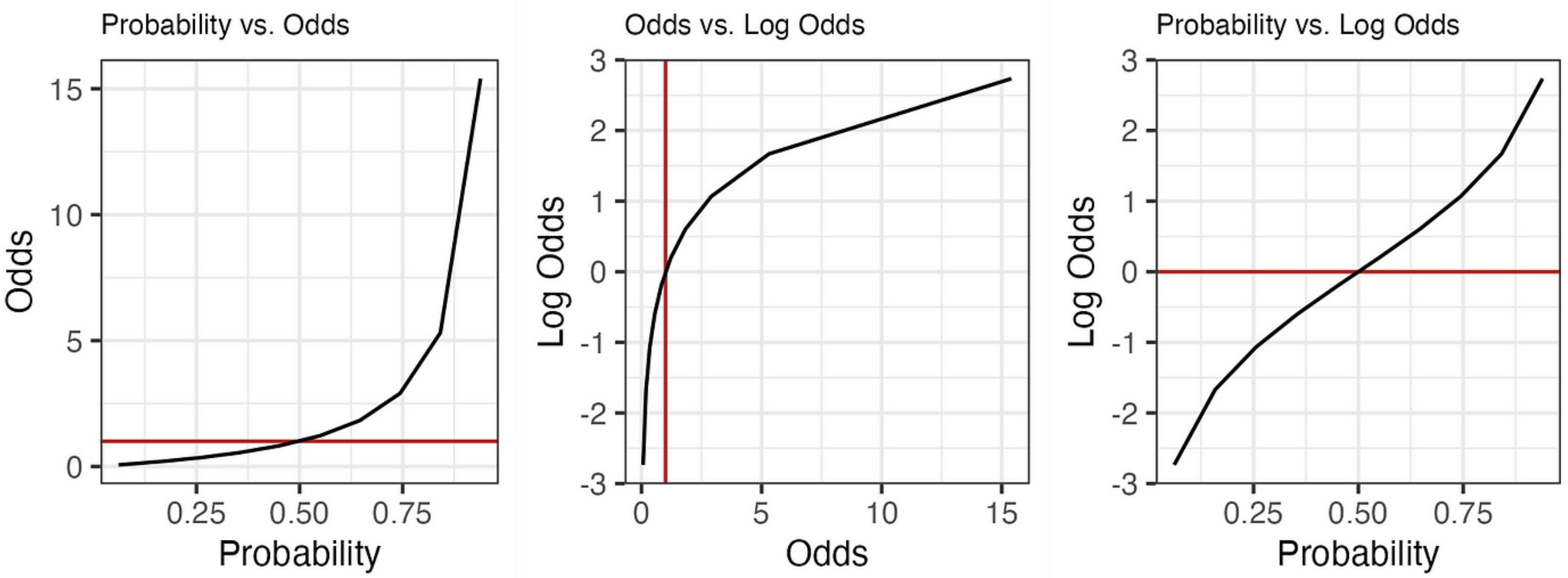
  0.7

- Alternatively, we can say, the odds ratio for passing the course is odds ratio = **0.7/0.3 = 2.3 or 2.3: 1**. Meaning, the chance of passing is 2.3 times that of failing. **(Comparatively)**

- log odds is just the logarithm of odd ratio : log(2.3)

Connecting Probability, Odds, and Log Odds

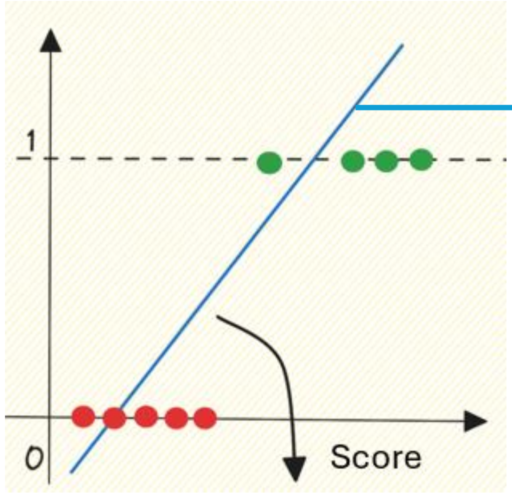
$$odds = \frac{p}{(1-p)} \qquad \log(odds) = log\left(\frac{p}{1-p}\right)$$



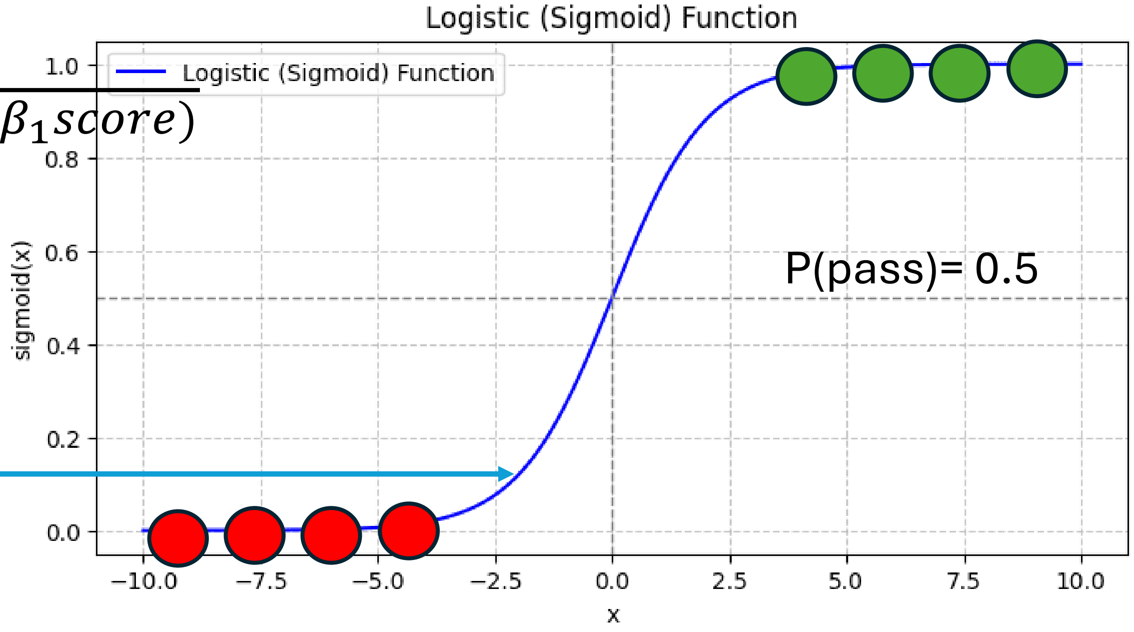
Connecting Probability, Odds, and Log Odds

$$\frac{1}{1+e^{-(\beta_0+\beta_1 score)}}$$

$\beta_0 + \beta_1 score$



P(pass)



$\beta_0 + \beta_1 score$

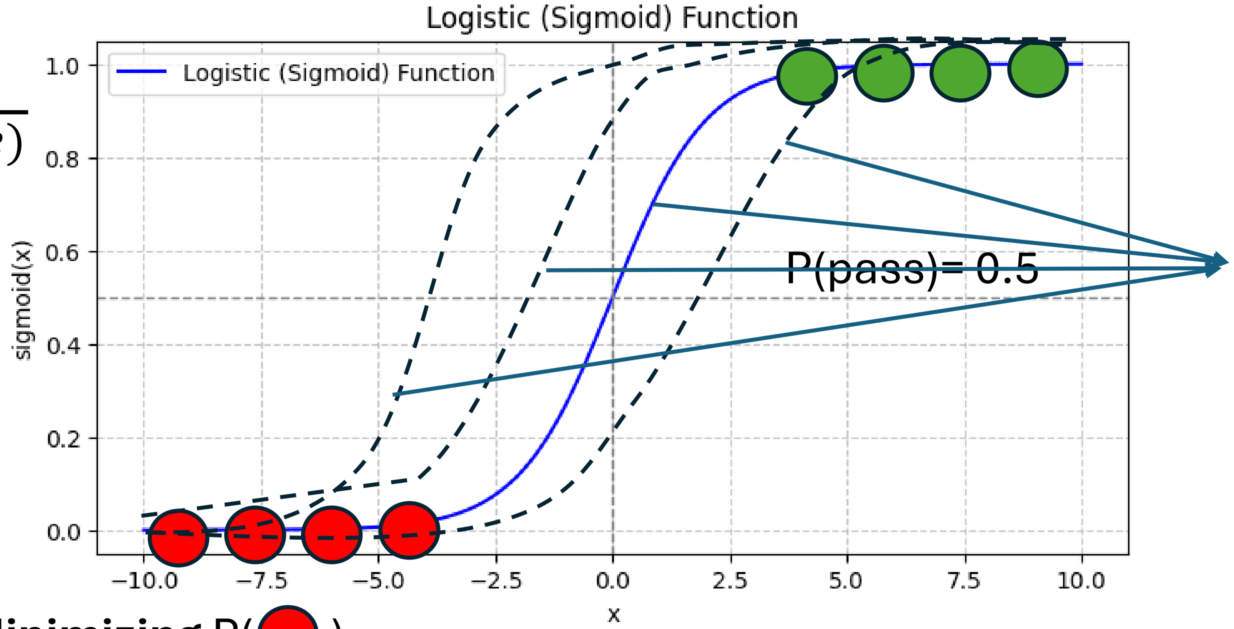
- $\beta_0 + \beta_1 score$ (-inf, inf) is log odds
- $e^{-(\beta_0+\beta_1 score)}$ (0, inf) is odds
- $\frac{1}{1+e^{-(\beta_0+\beta_1 score)}}$ (0,1) is probability

Model Training/Fitting

Maximizing P( )

$$\frac{1}{1+e^{-(\beta_0+\beta_1 score)}}$$

P(pass)



Parameterized by

β_0, β_1

Minimizing P( )

$\beta_0 + \beta_1 score$

Or **maximizing** 1 - P( )

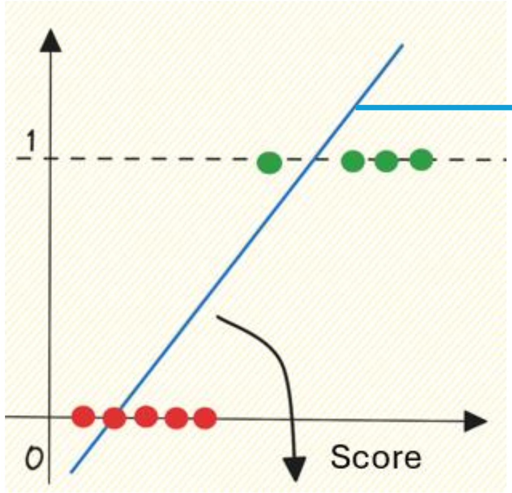
Maximum Likelihood Estimate

$$L(\beta) = \prod_{s \ in \ y_i = 1} p(x_i) \ * \prod_{s \ in \ y_i = 0} (1 - p(x_i))$$

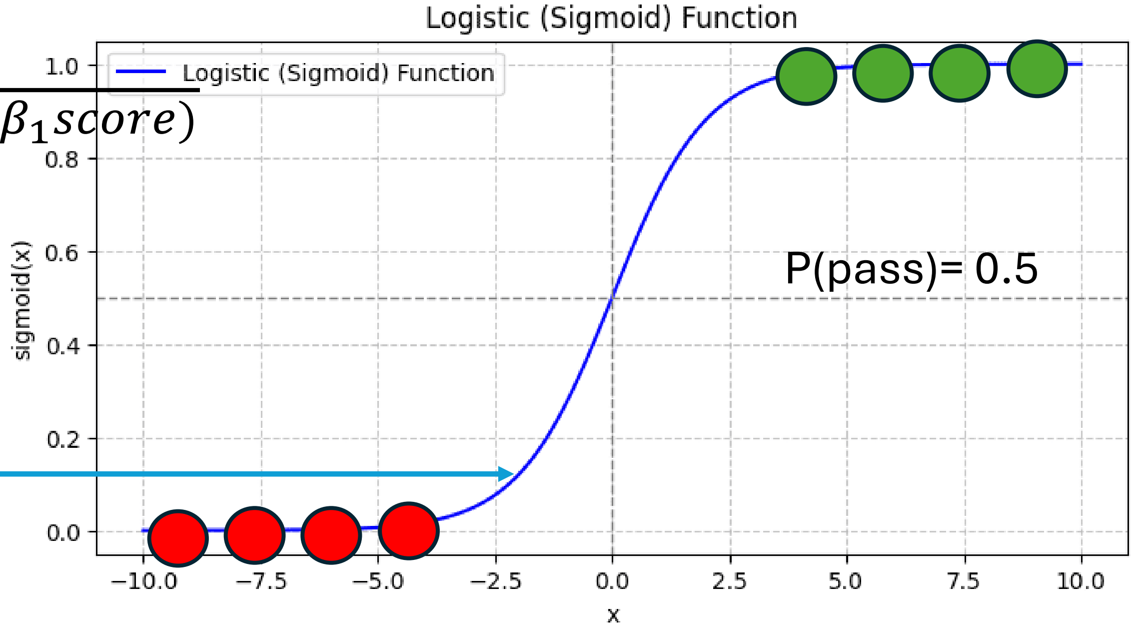
Interpreting Coefficients

$$\frac{1}{1+e^{-(\beta_0+\beta_1 score)}}$$

$\beta_0 + \beta_1 score$



P(pass)



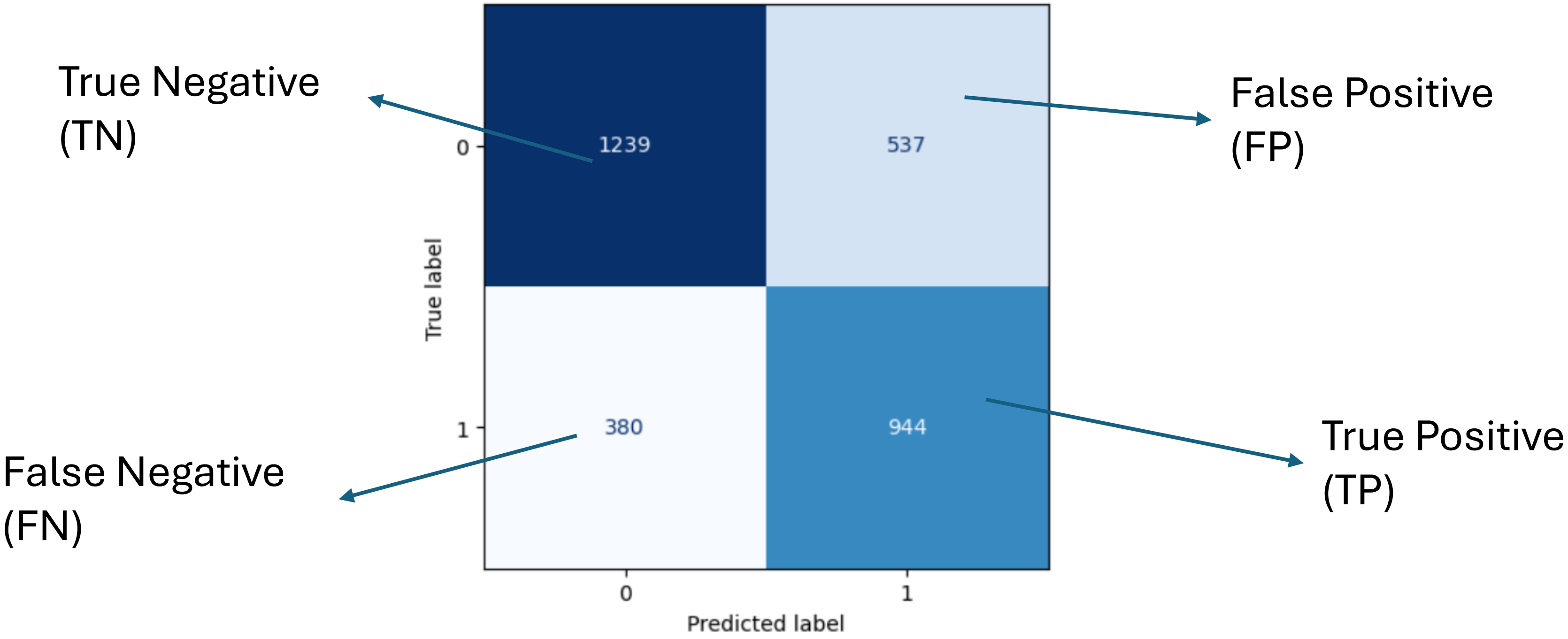
$\beta_0 + \beta_1 score$

- $\beta_0 + \beta_1 score$ (-inf, inf) is log odds
- $e^{-(\beta_0+\beta_1 score)}$ (0, inf) is odds
- $\frac{1}{1+e^{-(\beta_0+\beta_1 score)}}$ (0,1) is probability

One unit change in $score$ results in β_1 change in log odds and $e^{-\log odds}$ in odds ratio.

Evaluating a LR Model – Confusion Matrix

- Confusion Matrix: A table used to describe the performance of a classification model. It presents the number of true positives, true negatives, false positives, and false negatives.



Evaluating a LR Model – Accuracy

- Accuracy is defined as the number of correct predictions over the total predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Evaluating a LR Model – Accuracy

- Precision: It is the proportion of true positive predictions out of all positive predictions made by the model. Precision is useful when the cost of false positives is high (e.g., fraud).

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

Evaluating a LR Model – Recall (Sensitivity)

- Recall: It is the proportion of true positive instances that were correctly identified by the model. Recall is useful when the cost of false negatives is high (e.g., Customer Churn Prediction).

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

Evaluating a LR Model – F1 Score

- Recall: It is the harmonic mean of precision and recall. It provides a balance between precision and recall. It is a useful metric when there is an uneven class distribution.

$$F1\,Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Practice