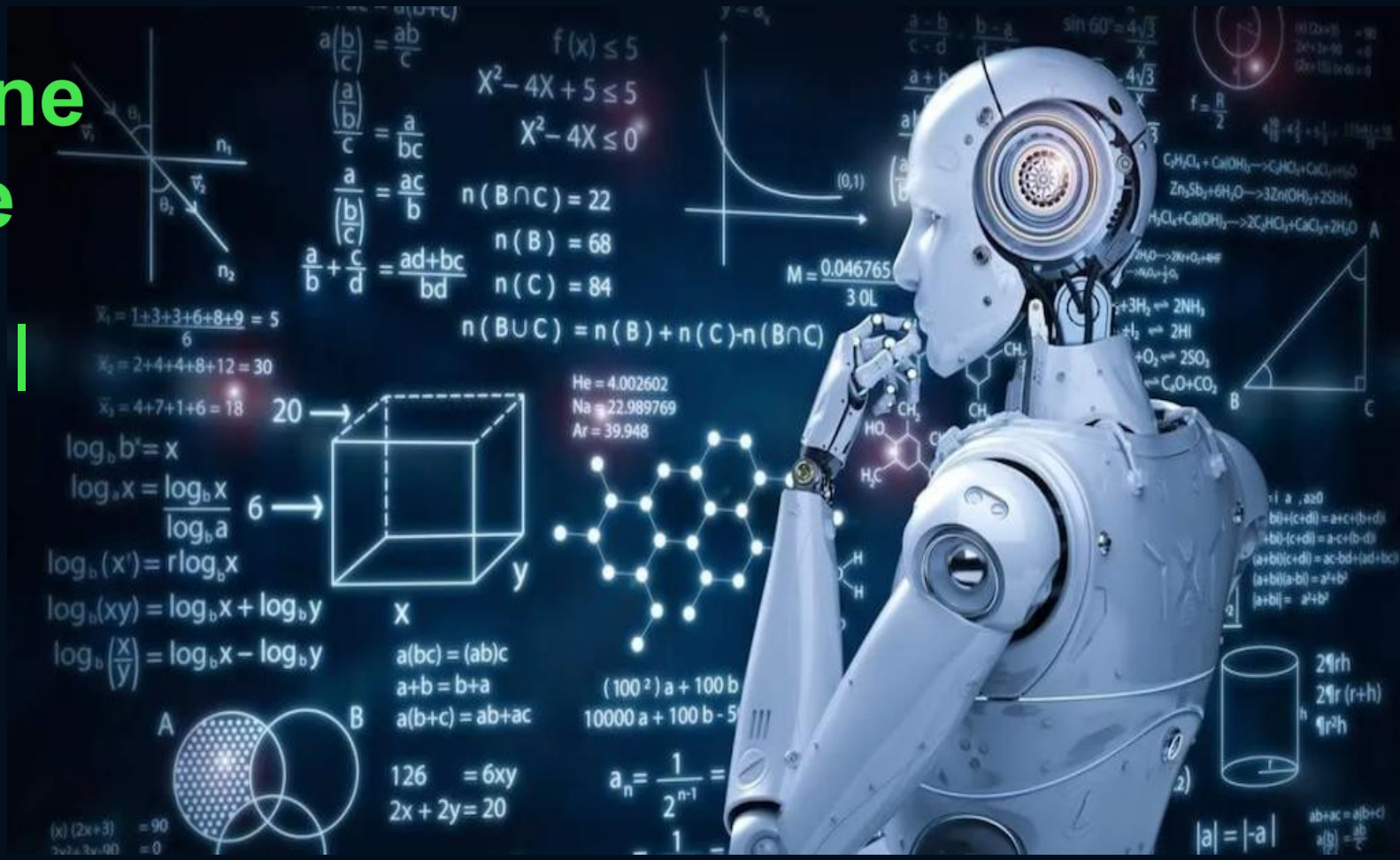
NBS Nottingham Business School

Principles of Machine Learning in Finance

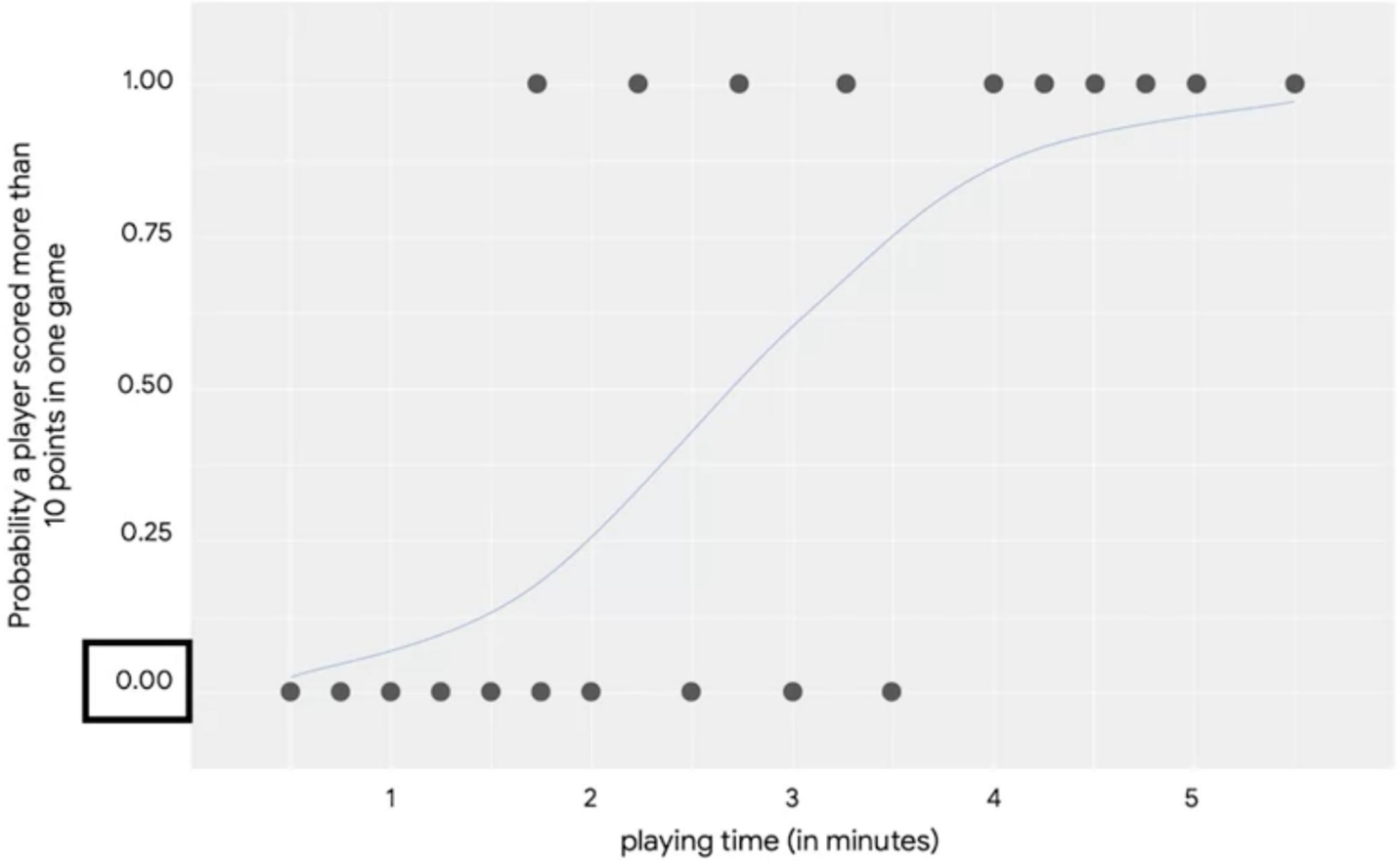
3. Supervised Learning | Classification | Logistic Regression



Learning Outcomes

- Logistic Regression and Binomial logistic regression
- Maximum likelihood estimation
- Logistic regression in Python
- Evaluating logistic regression
- Interpreting results in logistic regression
- **Coding Activity 3**: Supervised ML. Classification.
  Logistic Regression ||
  [Binomial Logistic Regression with Python]

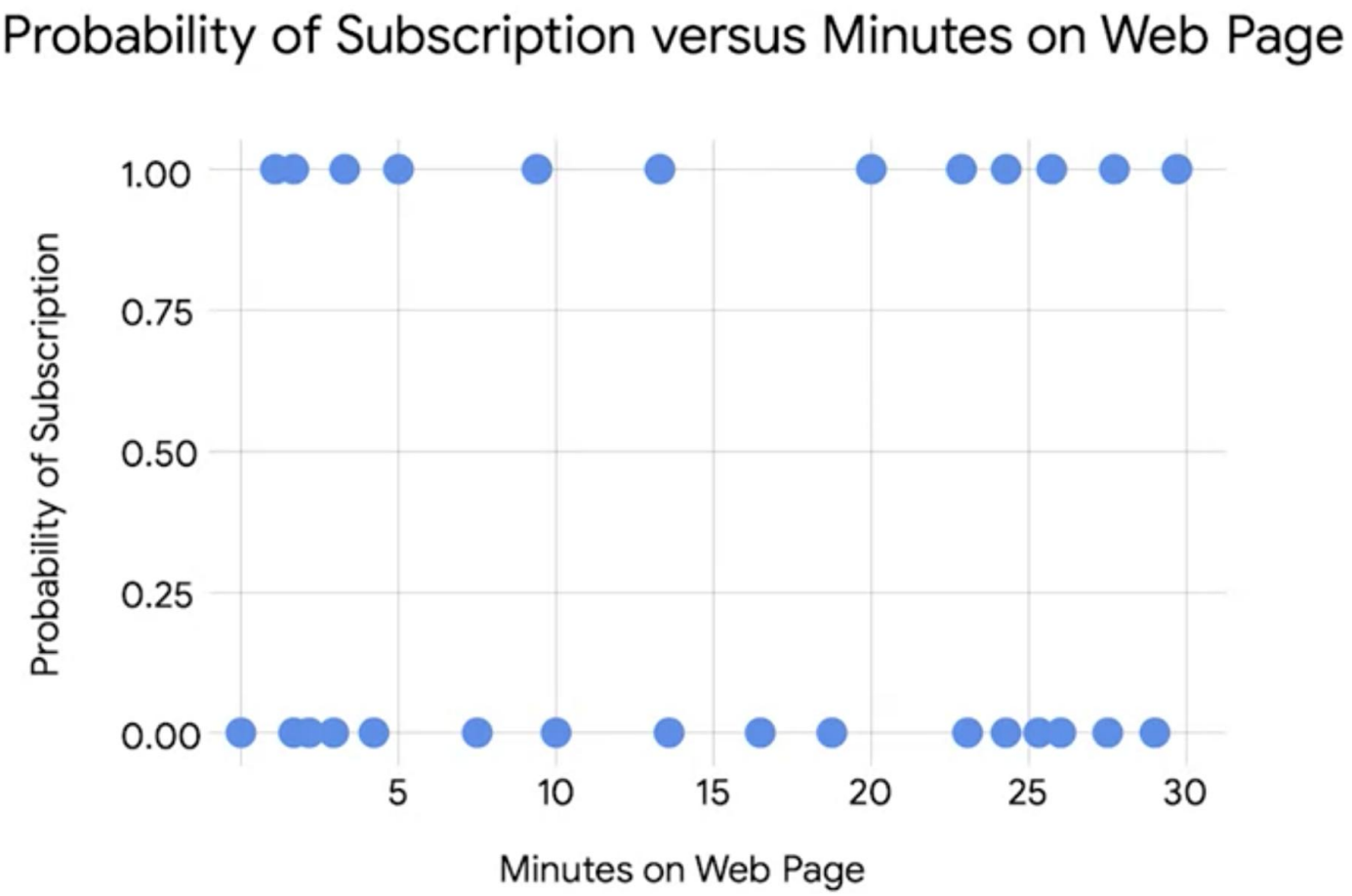
Logistic Regression



Logistic regression is a technique that models a categorical dependent variable (Y) based on one or more independent variables (X)

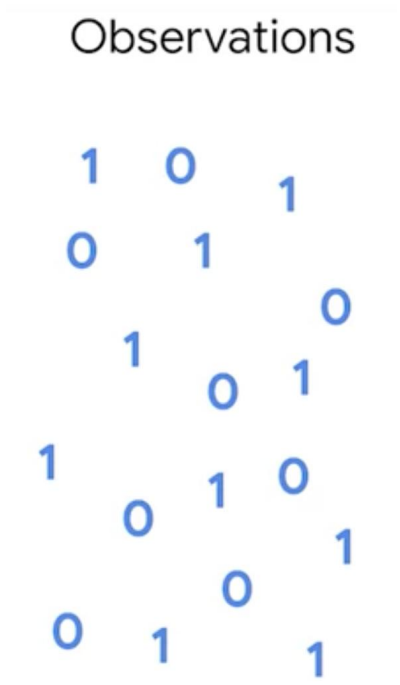
Example 1. Logistic Regression

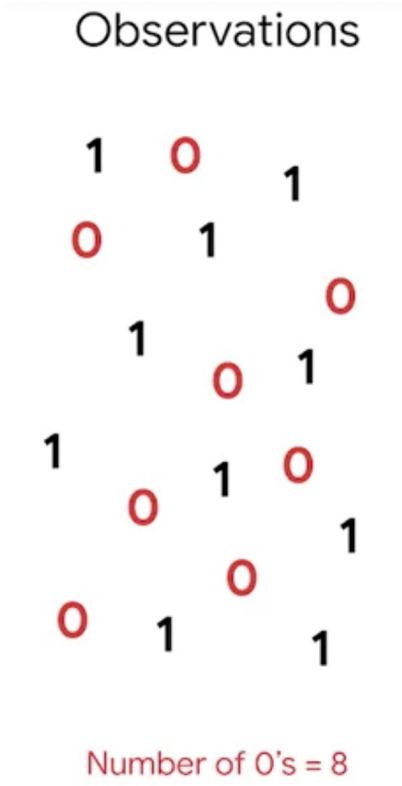
Y	X
Users don't subscribe (Y = 0) Users subscribe (Y = 1)	Continuous Minutes the user spends on a webpage

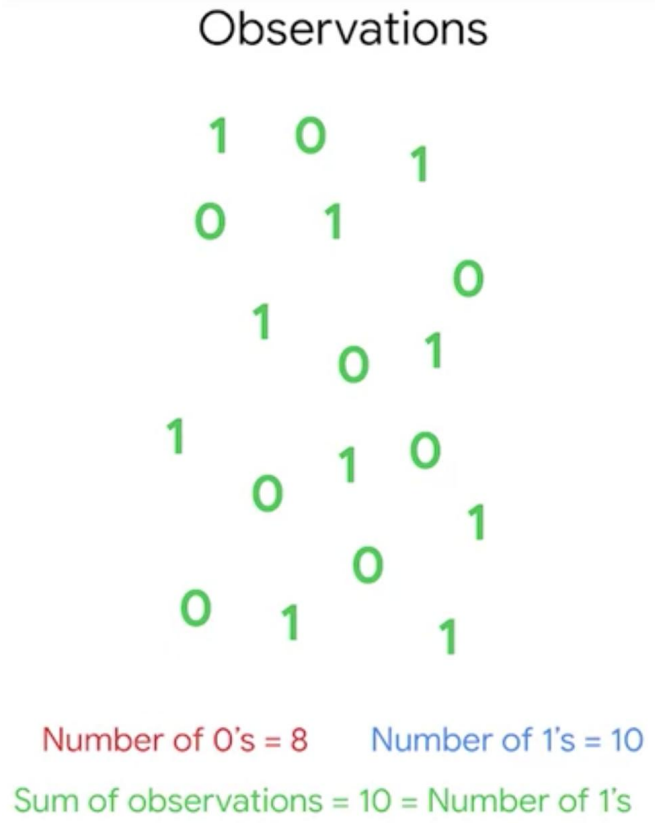


Logistic Regression Model

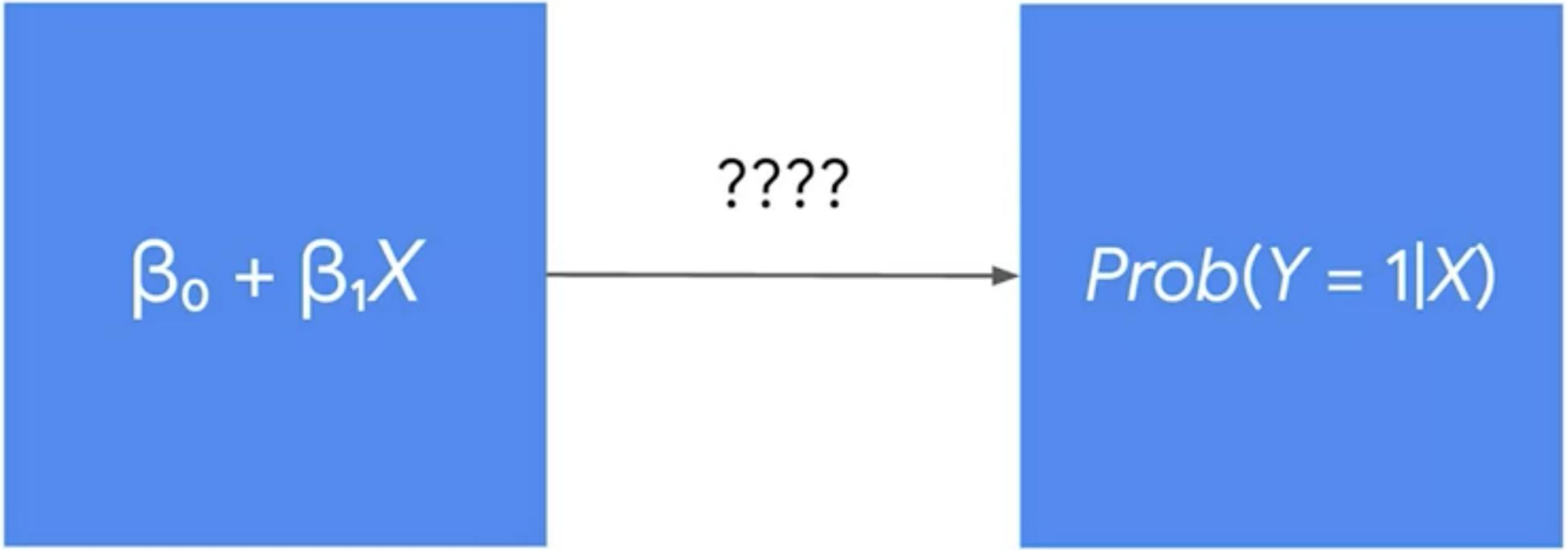
$$\mu\{Y|X\} = Prob(\text{Y=1|X}) = p$$







Logistic Regression Model (2)



Logistic Regression Model (3)

$$\mu\{Y|X\} = Prob(\text{Y=1|X}) = \text{p}$$

$$g(p) = \beta_0 + \beta_1 \bullet X$$

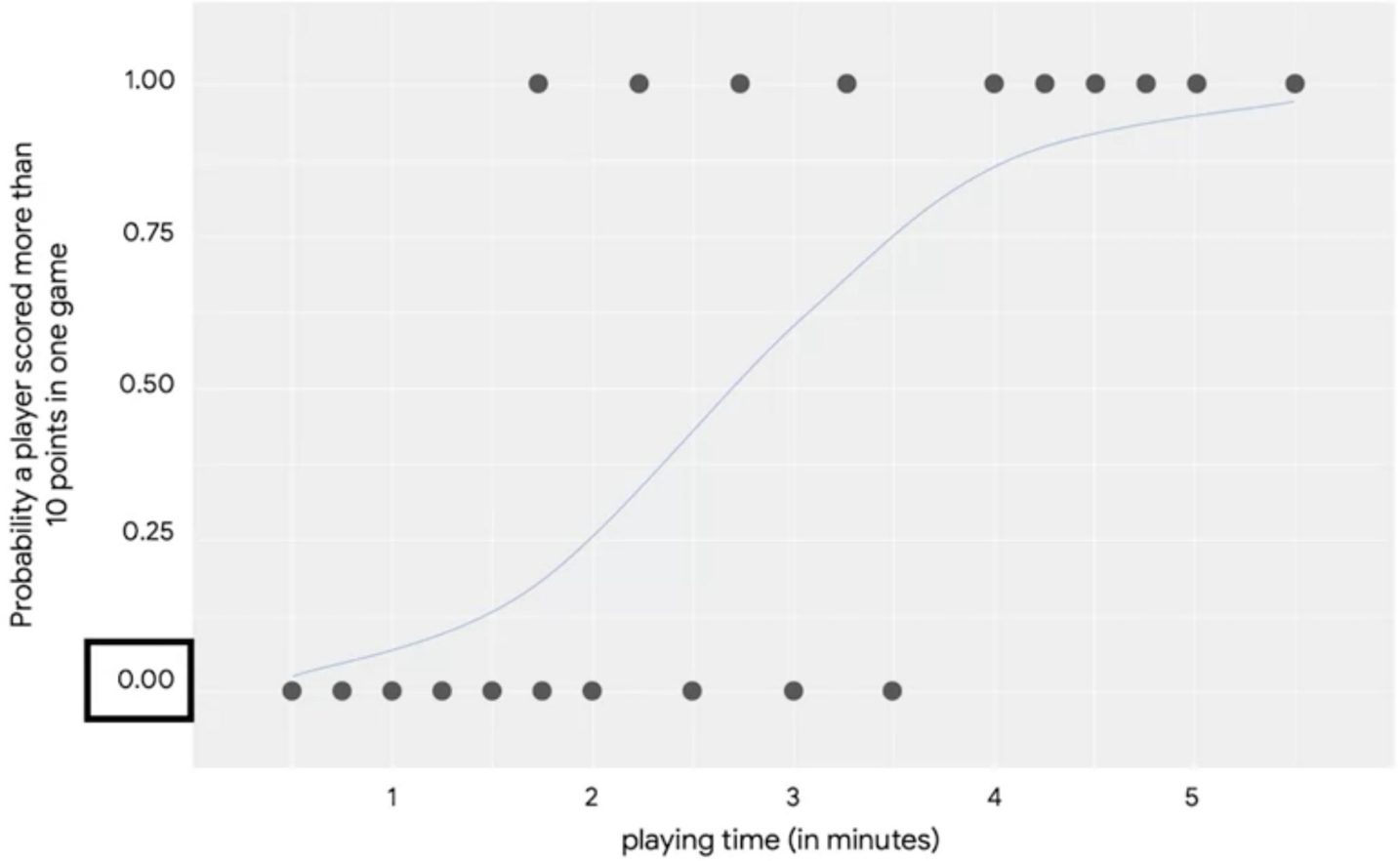
Link function

Link function is a non-linear function that connects or links the dependent variable to the independent variables mathematically

Logistic Regression vs Linear Regression

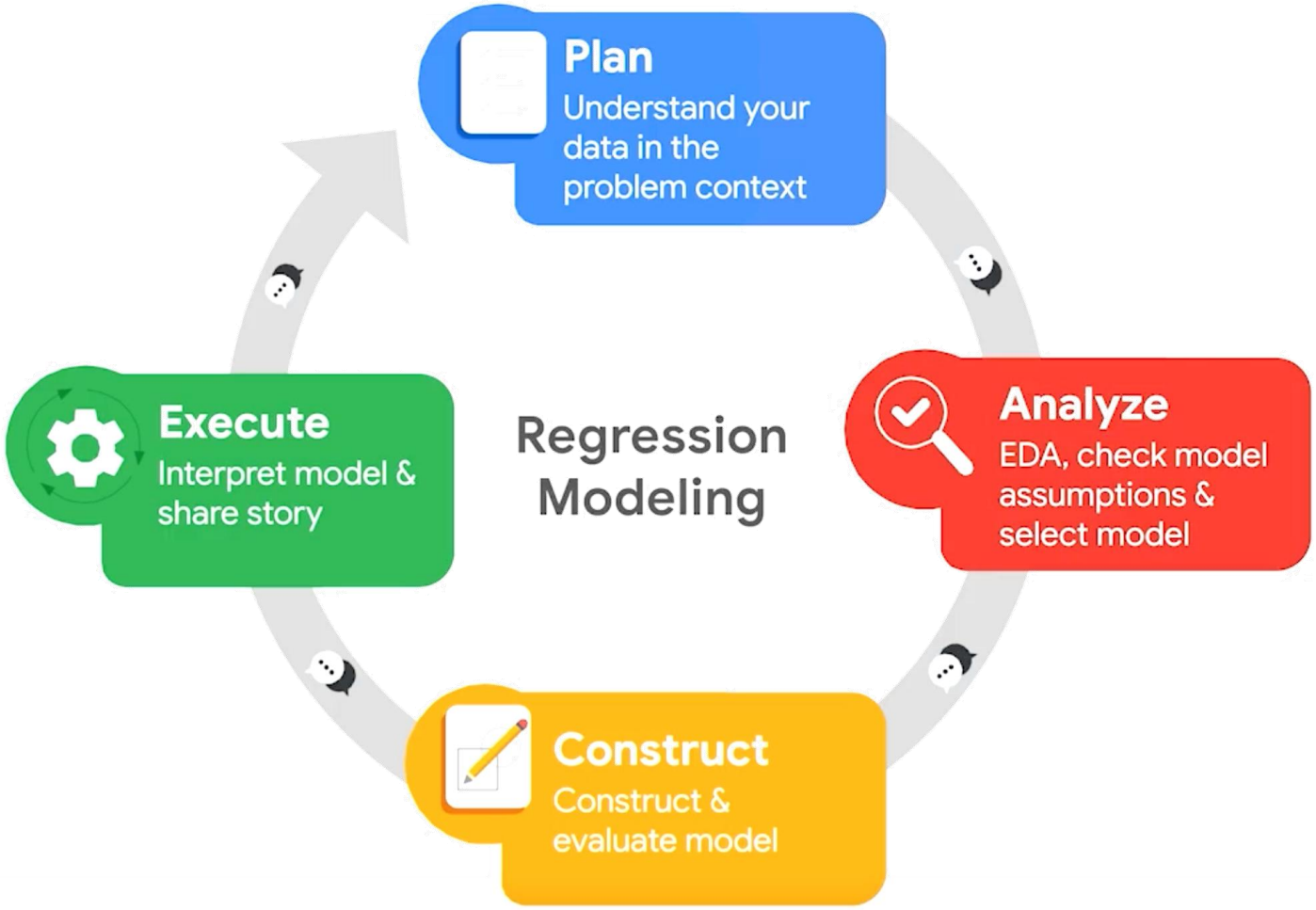
Linear Regression	Logistic Regression
Continuous data (i.e. book sales - 100 books, 200 books, 437 books, etc.)	Categorical data (i.e. newsletter subscription - yes/no)
Estimating the MEAN of y	Estimating the PROBABILITY of an outcome
$\mu(Y\|X) = \beta_0 + \beta_1 X$	$\mu(Y\|X) = Prob(Y = 1\|Y) = p$ $g(p) = \beta_0 + \beta_1 X$

Binomial Logistic Regression



Binomial logistic regression is a technique that models the probability of an observation falling into one of two categories, based on one or more independent variables

PACE in Regression Modeling



Binomial Logistic Regression: Assumptions

- **Linearity**: There should be a linear relationship between each X variable and the logit of the probability that Y equals 1

- **Independent observations**:

$$P(A \ AND \ B) = P(A) \bullet P(B)$$

- **No multicollinearity** between the independent variables
- **No extreme outliers** in the dataset

Logit (log-odds)

$$Odds = \frac{p}{1-p}$$

Example:

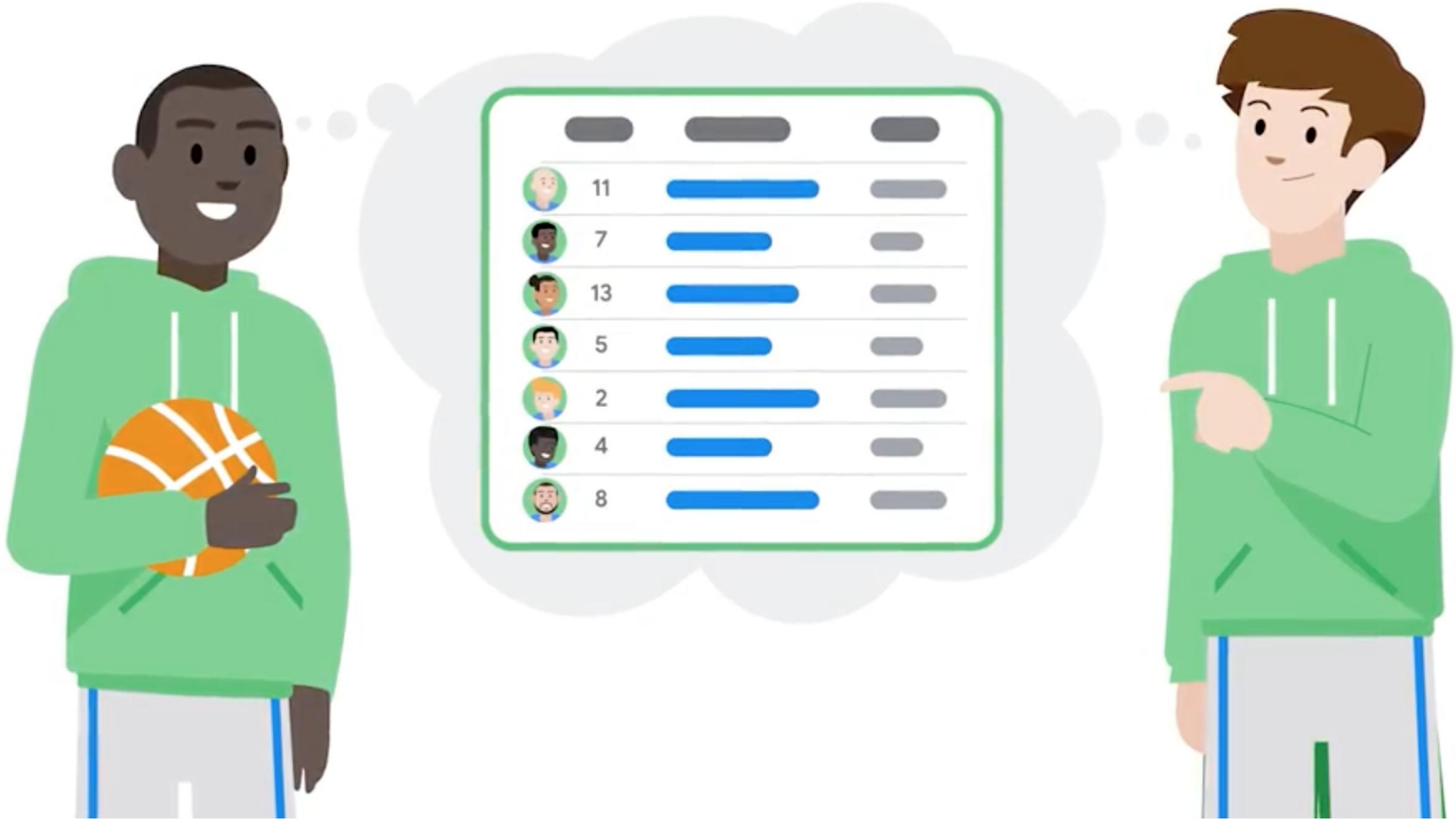
p =0.6; (1-p) = 0.4; Odds = 0.6 / 0.4 = 1.5

Logit (log-odds) is the logarithm of the odds of a given probability.

The logit of probability p is equal to the logarithm of p divided by 1 minus p:

$$logit(p) = log(\frac{p}{1-p})$$

Example 2. Likelihood of Players Scoring Many Points in the Game



Logit of p in terms of X variables

$$logit(p) = log(\frac{p}{1-p})$$

$$logit(p) = \beta_0 + \beta_1 \bullet X_1 + \beta_2 \bullet X_2 + \ldots + \beta_N \bullet X_N$$

where:

N = number of independent variables

Maximum likelihood estimation (MLE)

Maximum likelihood estimation (MLE) is a technique for estimating the beta parameters that maximize the likelihood of the model producing the observed data.

Likelihood is the probability of observing the actual data, given some set of beta parameters.

Binomial Logistic Regression: Assumptions Recap

- **Linearity**: There should be a linear relationship between each X variable and the logit of the probability that Y equals 1

- **Independent observations**:

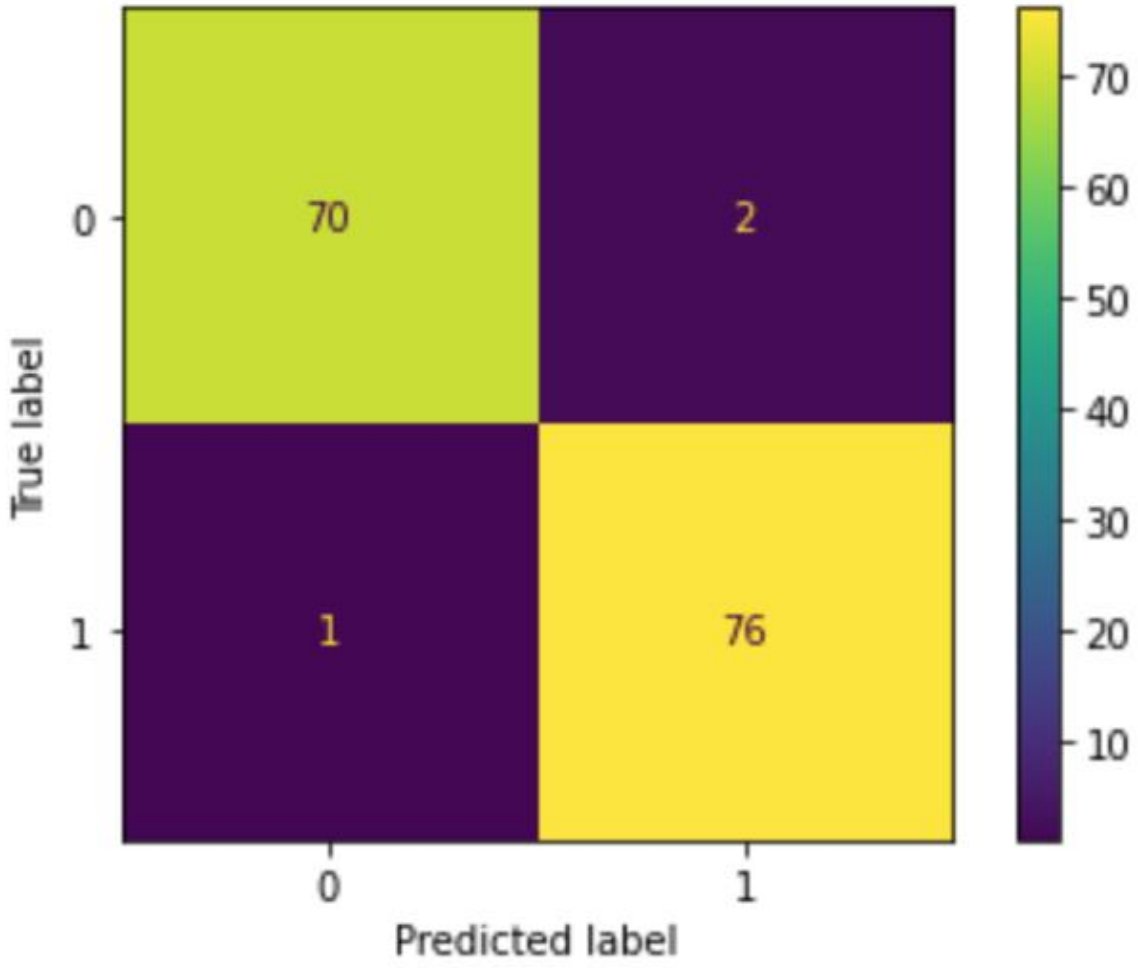
$$P(A \ AND \ B) = P(A) \bullet P(B)$$

- **No multicollinearity** between the independent variables
- **No extreme outliers** in the dataset

The Best Logistic Regression Model

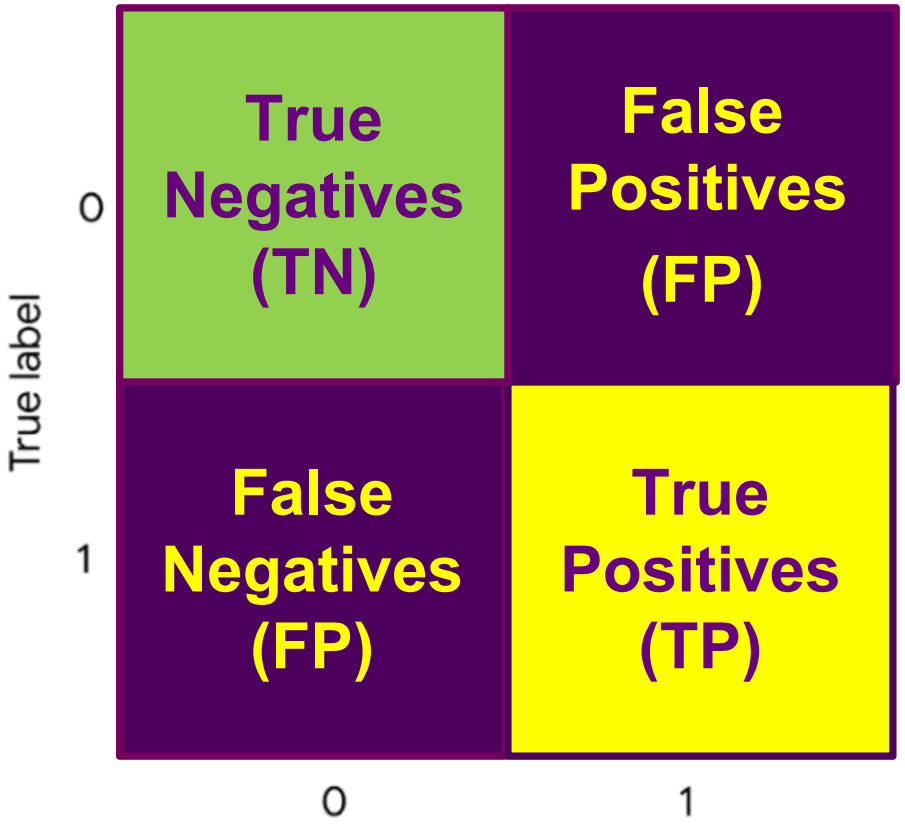
The best logistic regression model estimates the set of beta coefficients that maximizes the likelihood of observing all of the sample data.

Confusion Matrix



Confusion matrix is a graphical representation of how accurate a classifier is at predicting the labels for a categorical variable

Confusion Matrix (2)



- **True negatives**: The count of observations that a classifier correctly predicted as False (0)
- **True positives**: The count of observations that a classifier correctly predicted as True (1)
- **False positives**: The count of observations that a classifier incorrectly predicted as True (1)
- **False negatives**: The count of observations that a classifier incorrectly predicted as False (0)

Logistic Regression: Evaluation Metrics

I. The proportion of positive predictions that were true positives

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

II. The proportion of positives the model was to able identify correctly

$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

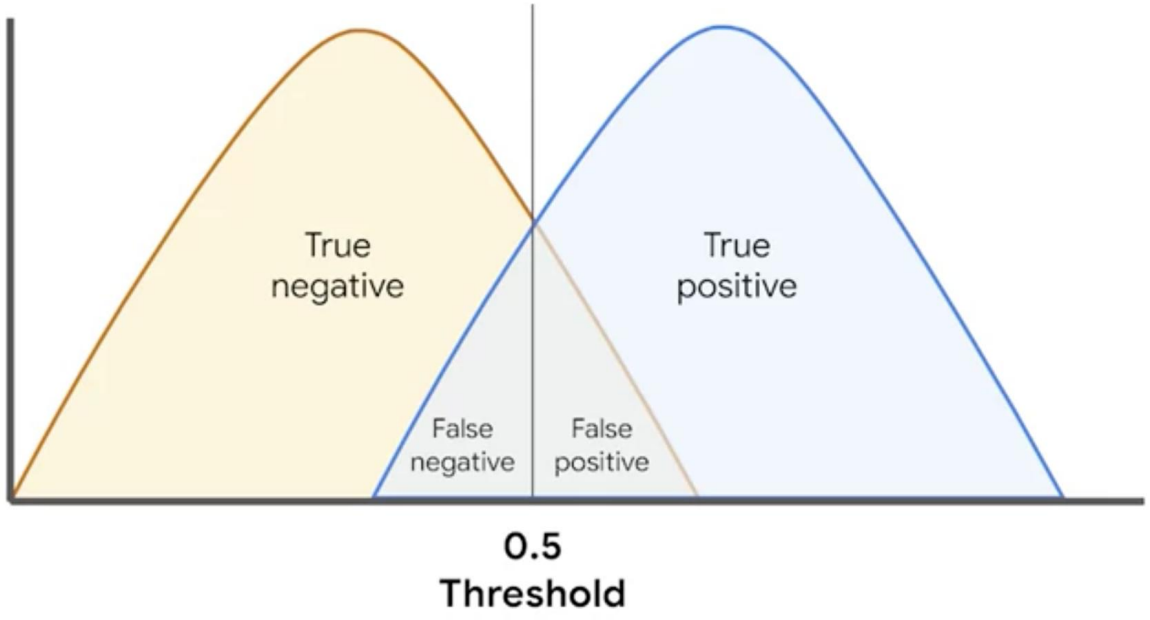
III. The proportion of data points that were correctly categorized

$$\textbf{Accuracy} = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

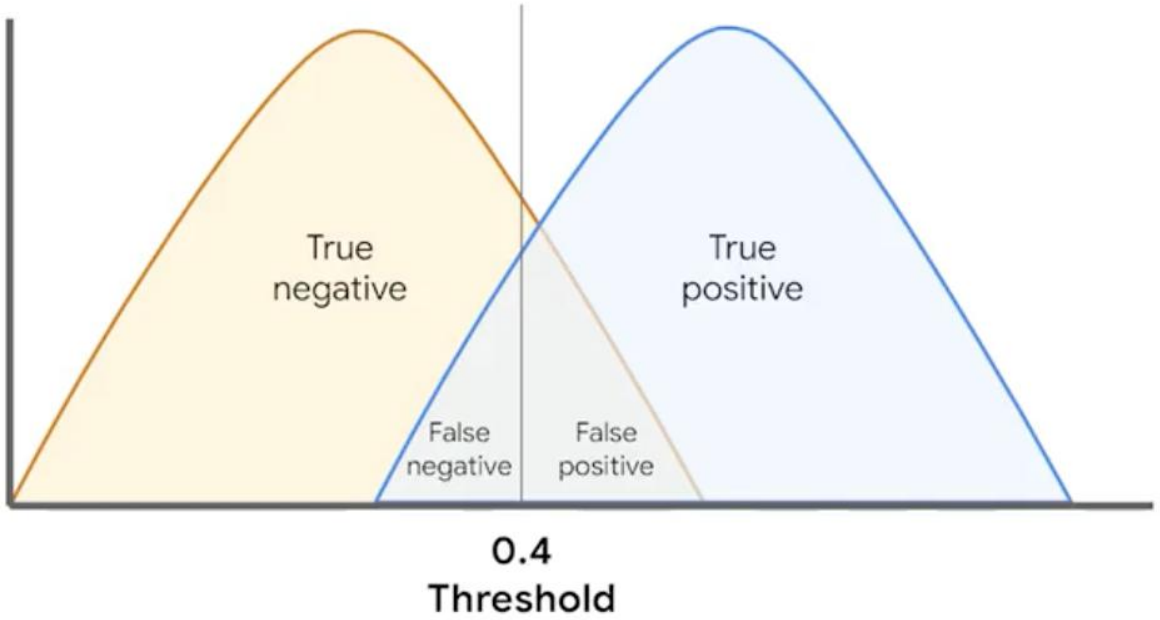
ROC Curve

ROC curve helps in visualizing the performance of a logistic regression classifier. ROC curve stands for receiver operating characteristic curve

a). Threshold = 0.5

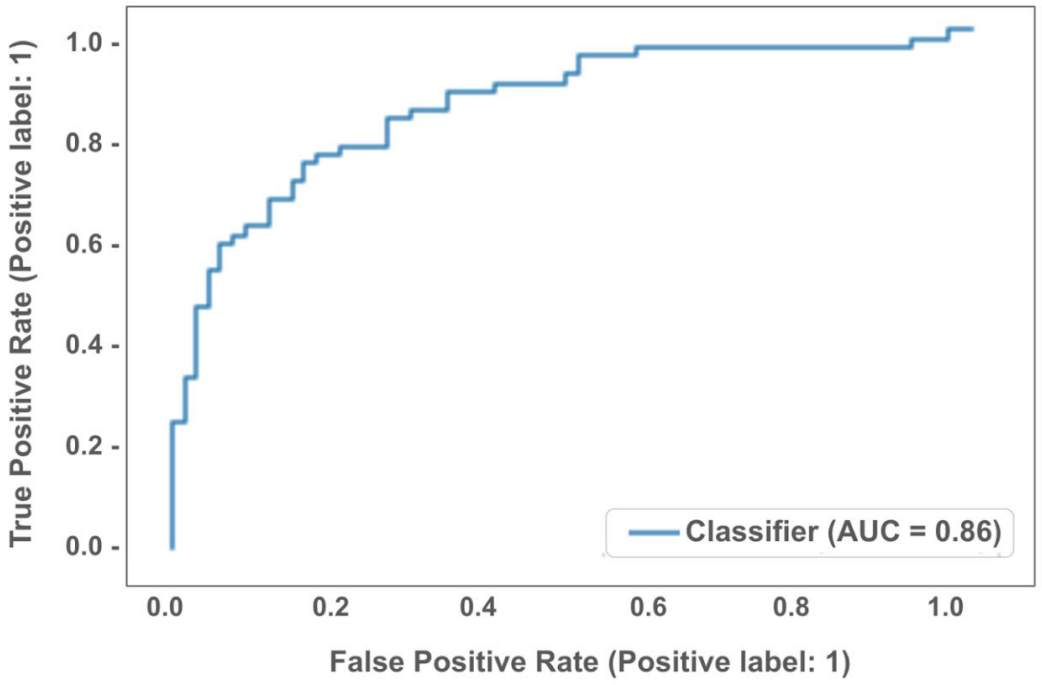


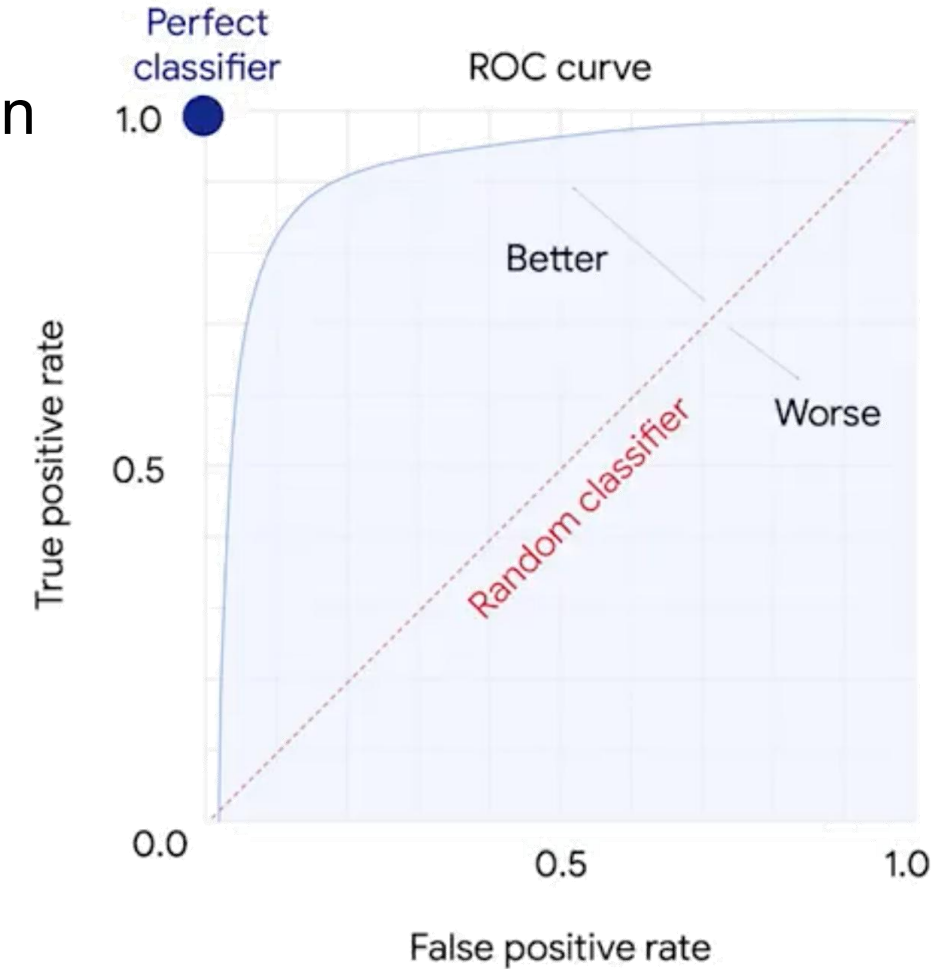
b). Threshold = 0.4



ROC Curve and AUC

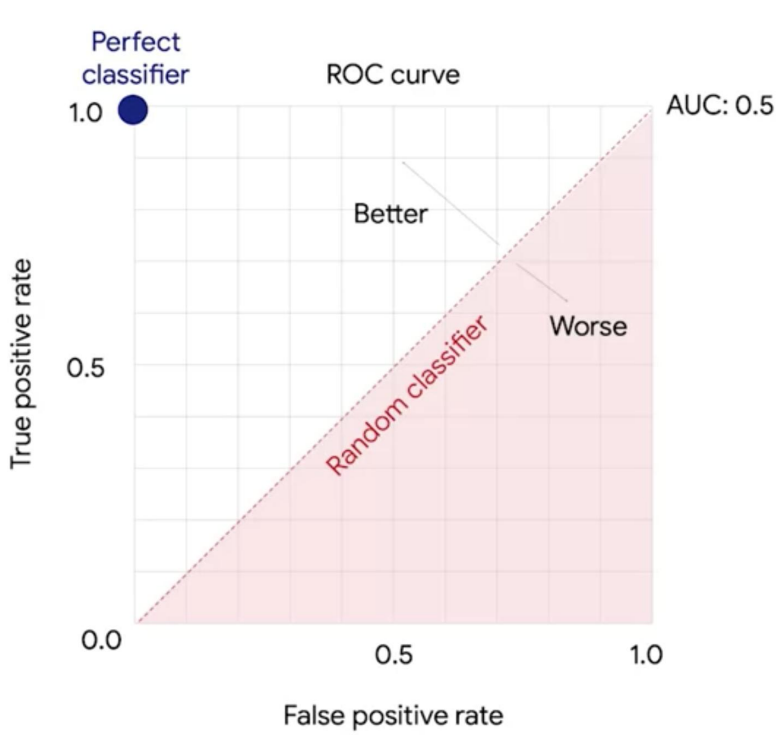
AUC provides an aggregate measure of performance across all possible classification thresholds.

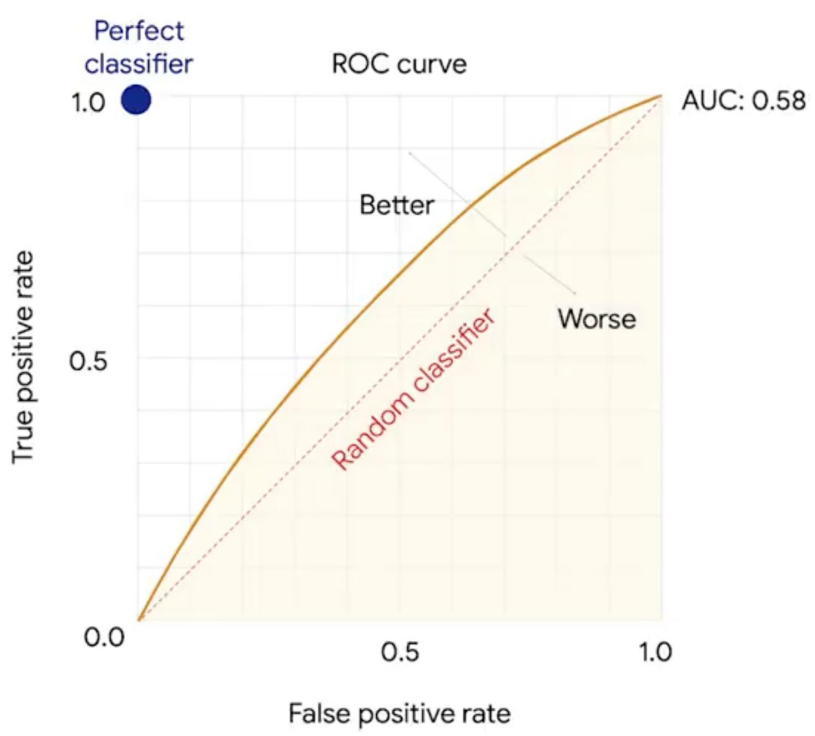


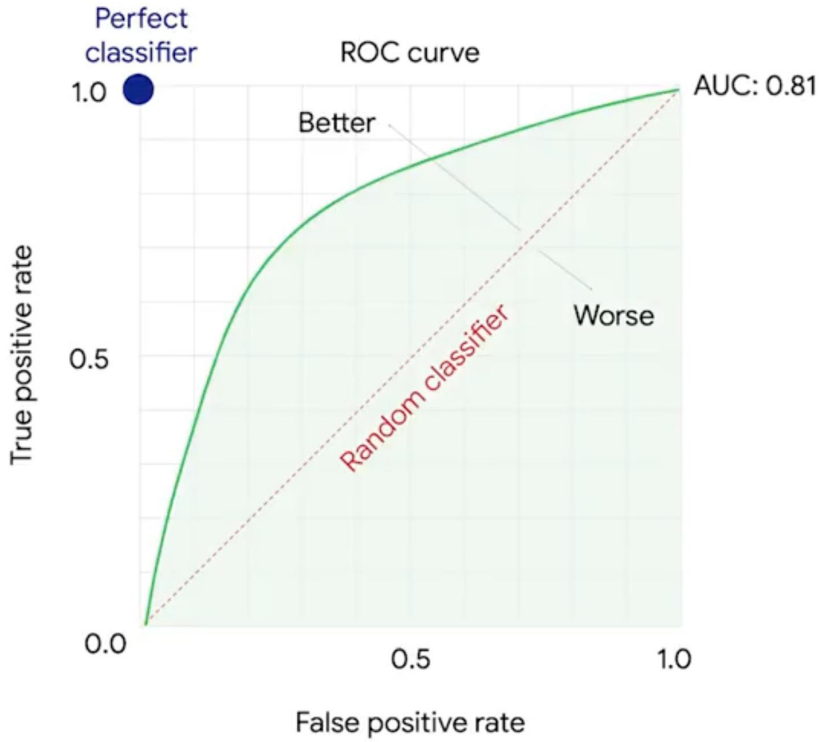


AUC: 0.96

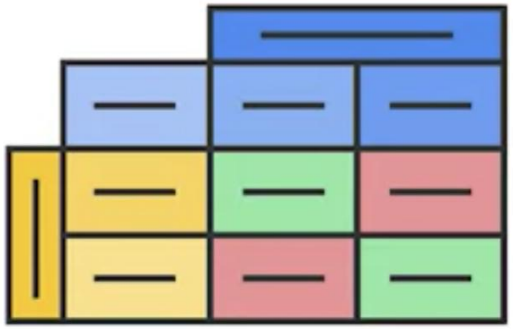
ROC Curve in AUC







Presenting Results



Confusion Matrix

0.85

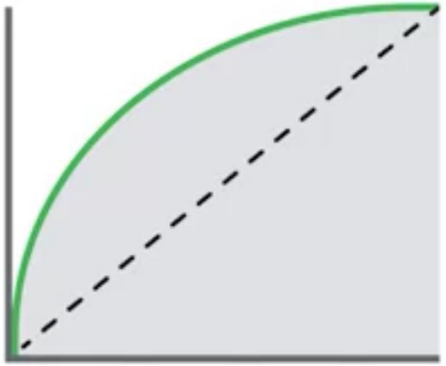
Accuracy

0.818

Precision

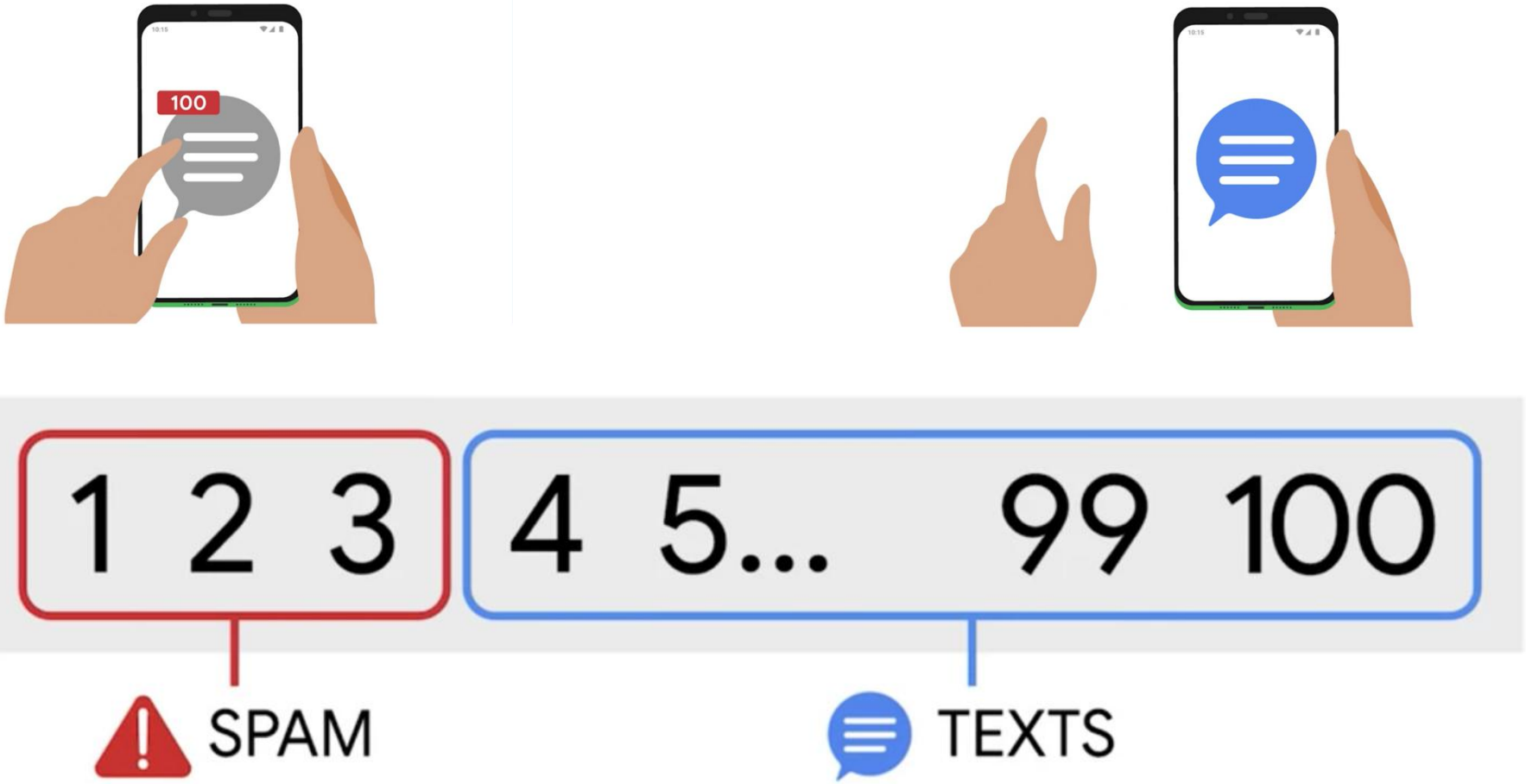
0.9

Recall



ROC/AUC

Example 3. Spam Detection



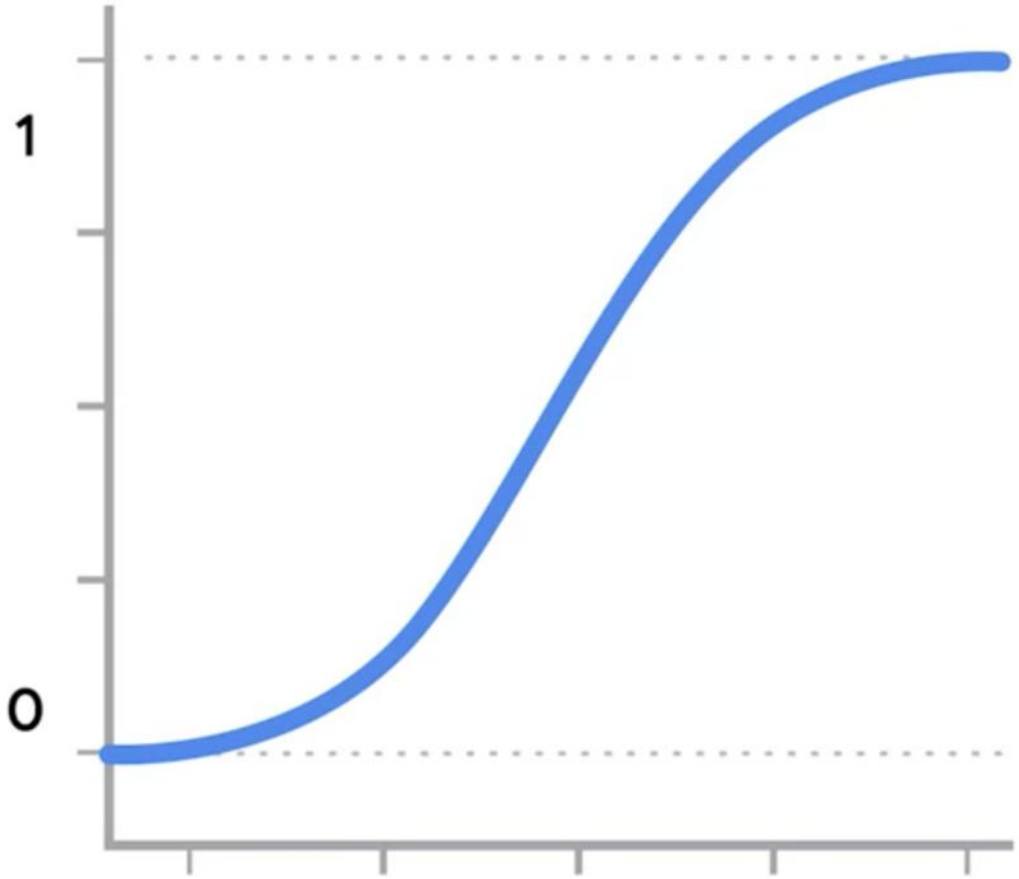
Example 3. Spam Detection (2)

0.25 NOT SPAM

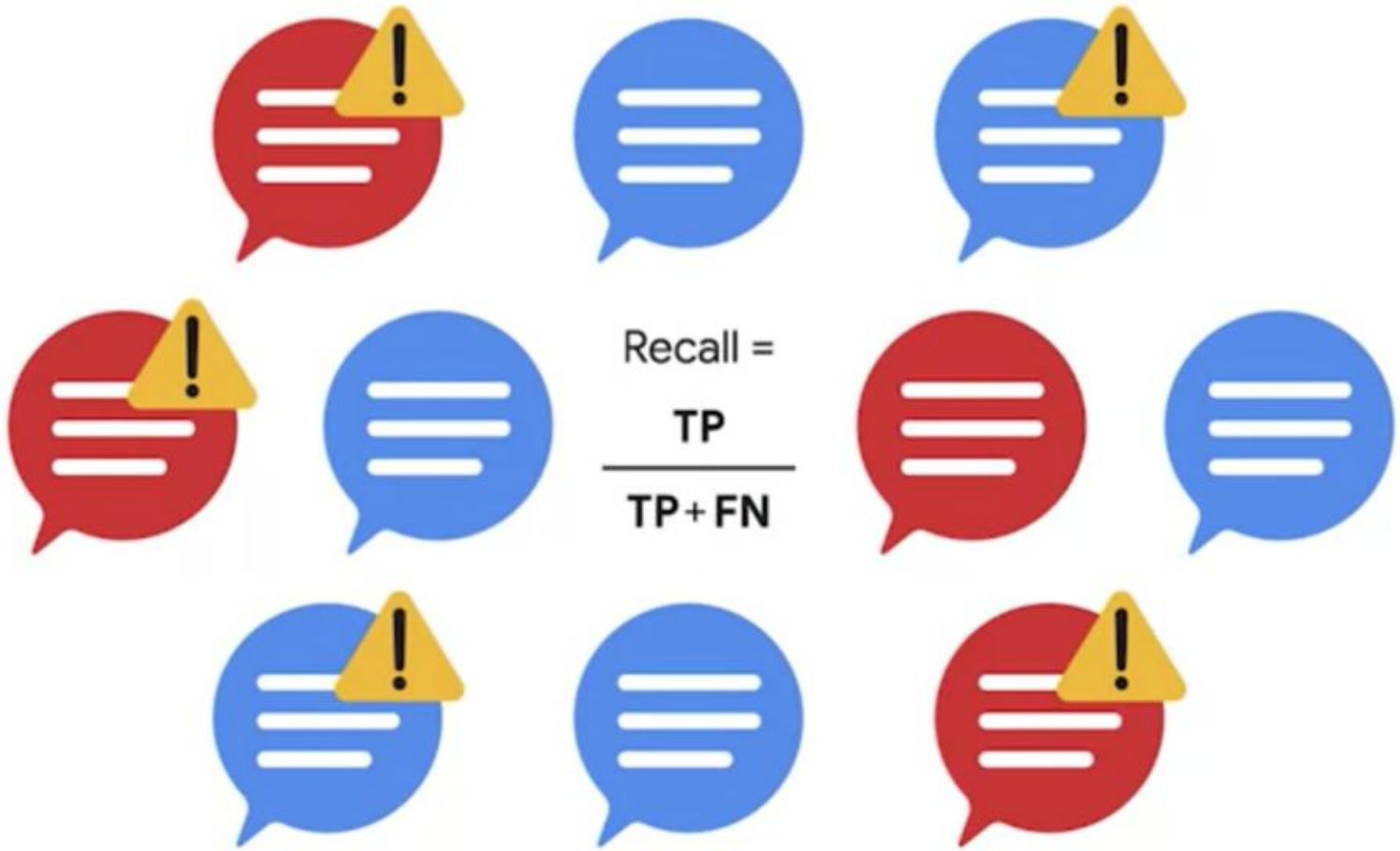
0.3 NOT SPAM

0.1 NOT SPAM

Accuracy = 97%

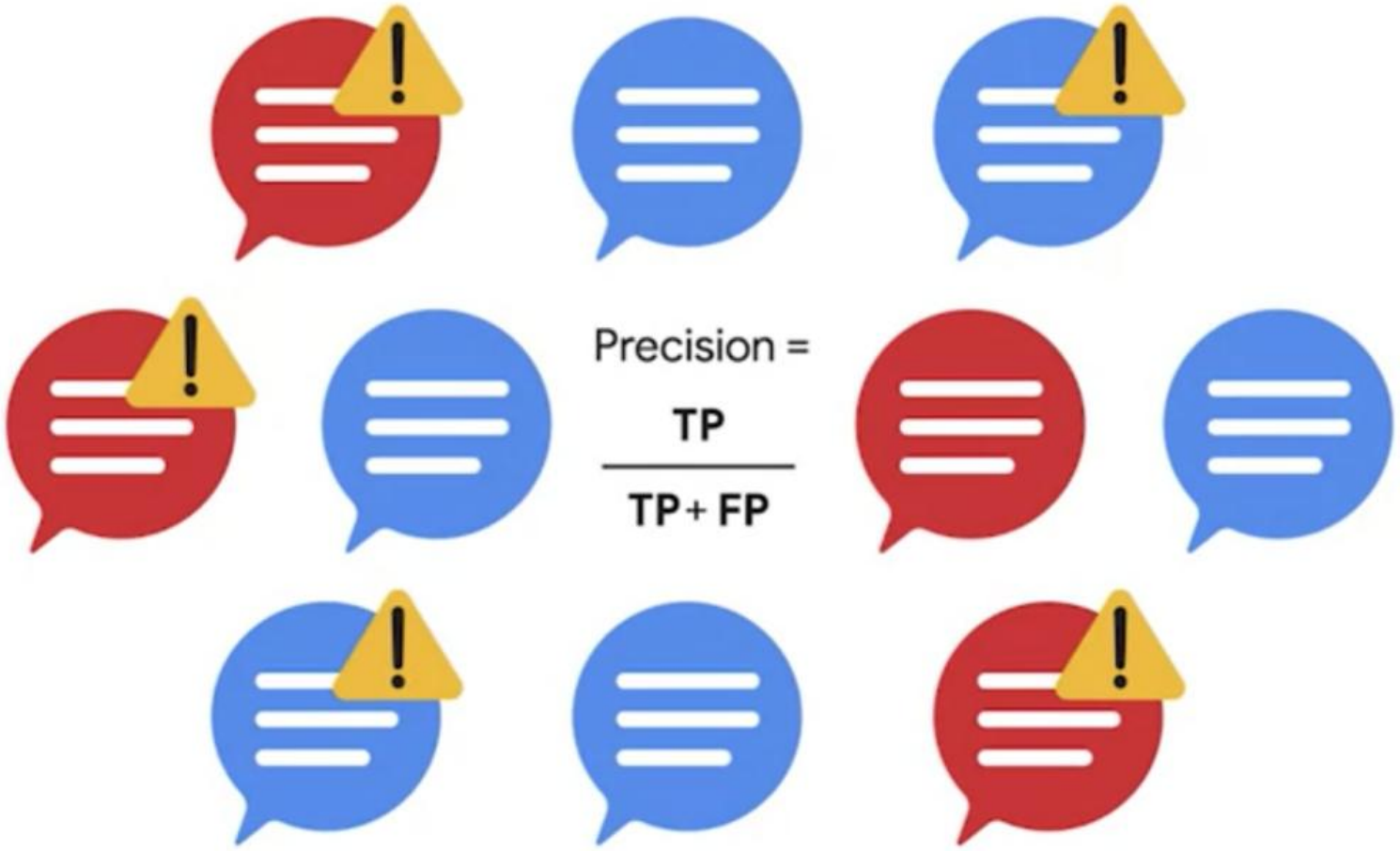


Example 3. Spam Detection (3)



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

Example 3. Spam Detection (4)



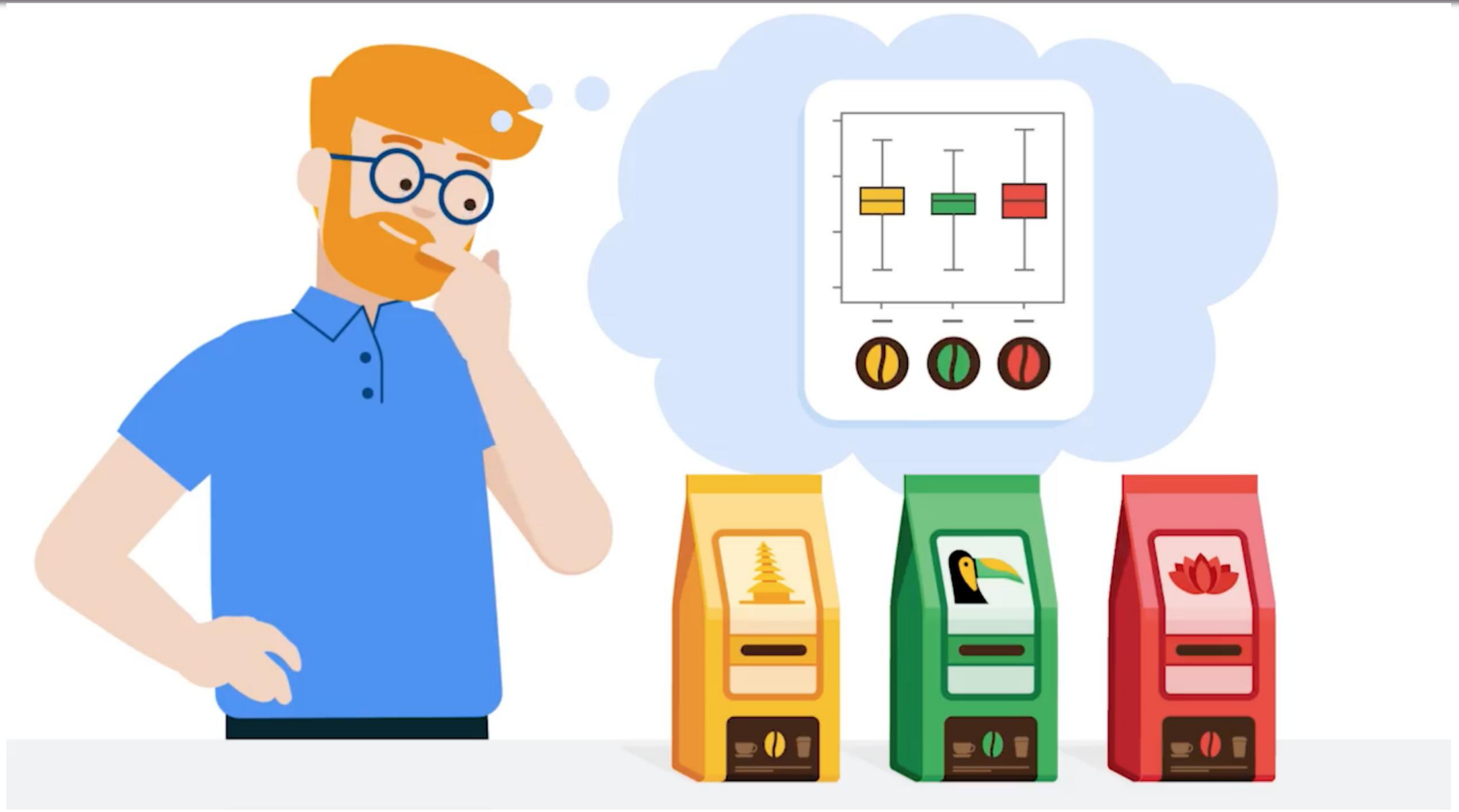
$$\text{Precision} = \frac{TP}{TP + FP}$$

Logit Regression: Revision

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 \bullet X_1$$

- A one-unit increase in vertical acceleration is associated with a β_1 increase in the log odds of p;
- e^{β_1} is how many times the odds of p will increase or decrease for every one-unit increase in vertical acceleration.

Example 4. Sampling Products for Better Sales



Example 4. Sampling Products for Better Sales (2)

Linear regression:

- Accessible interpretation
- Explain which factors impact the outcome variables
- Check model assumptions

Hypothesis test:

- Outcome variable is continuous
- Focus on comparing different groups

Example 4. Sampling Products for Better Sales (3)

Hypothesis test:

Null hypothesis (H_0): Approximately the same level of sales for each type of products

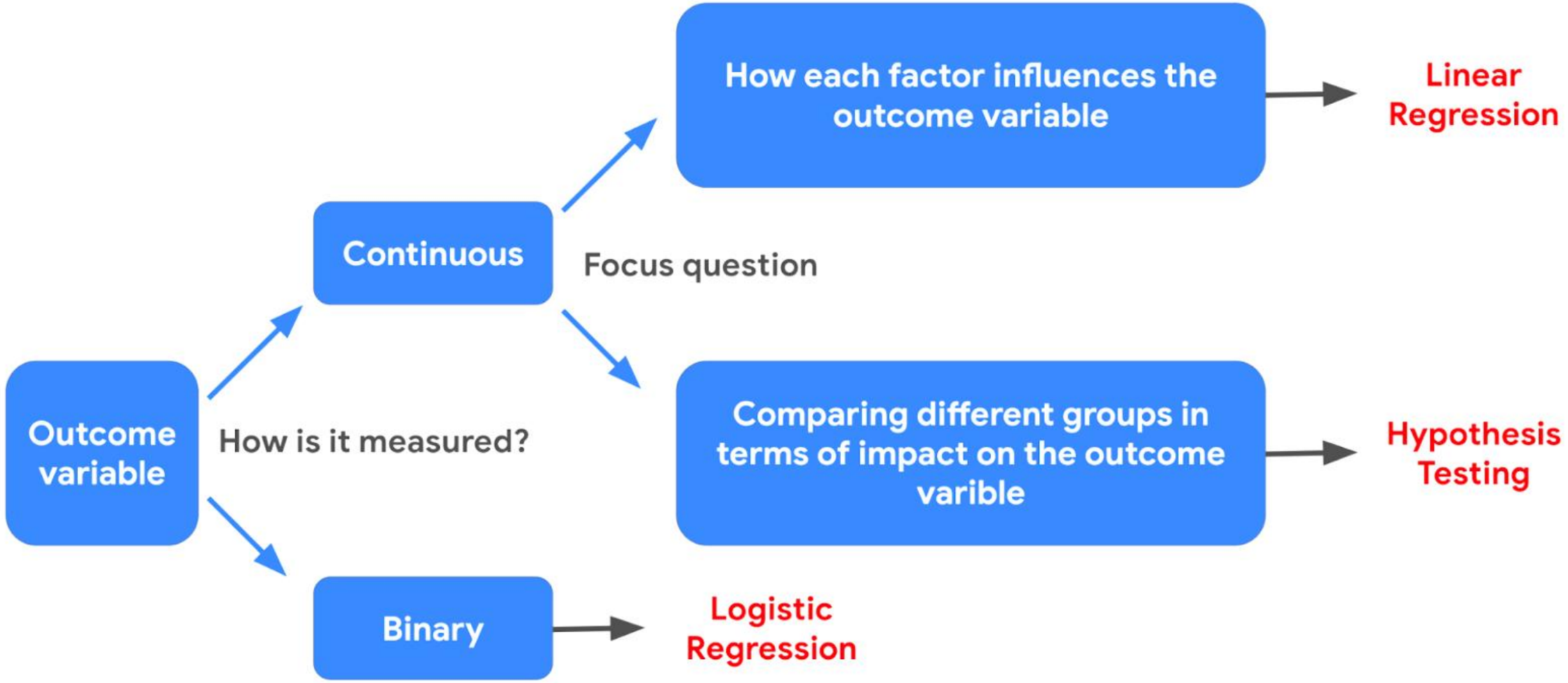
Alternative hypothesis (H_1): Not the same (different) level of sales for each type of products

Evaluation Logistic Regression: Revision

- P - value
- Confusion matrices
- Precision
- Recall
- Accuracy
- ROC/AUC
- AIC
- BIC

Logistic regression coefficients report in percentages how much a factor increases or decreases the **likelihood of an outcome**

Regression Models: Linear vs Logistic Regression



Coding Activity 3. Supervised ML. Binomial Logistic Regression

Lab 3. Supervised Machine Learning. Logistic Regression. Classification || Binomial Logistic Regression with Python

Steps to follow:

1. Upload the following files from the module learning room:
   - Jupiter notebook "Lab3_Logistic_Regression_with_Python.ipynb"
   - Csv-dataset file "activity.csv"
2. Follow along in the Jupiter notebook

Coding Activity 3. Interpreting Results

$$\beta_1 = -0.118$$

$$e^{\beta_1} = e^{-0.118} = 0.89$$

For every one-unit increase in the X_1, holding other variables constant, we expect that the odds Y being 1 to decrease by 11%

$$\beta_1 = 0.25$$

$$e^{\beta_1} = e^{0.25} = 1.28$$

For every one-unit increase in the X_1, holding other variables constant, we expect that the odds Y being 1 to increase by 28%

Q&A

Thank you!