

# Which model for which problem?

D8A.ACADEMY/RESOURCES/ALGORITHM-CHOOSER

Predict a number, explain the drivers



**Linear regression**  
coefficients are the explanation

Yes/no, and you must justify it

**Logistic regression**  
the baseline in banking for a reason

Yes/no, accuracy is all that matters



**Gradient boosting**  
XGBoost/LightGBM on tabular data

Find groups, no labels exist



**k-means**  
then LOOK at the clusters

Too many columns to see anything



**PCA**  
compress first, model second

Forecast with trend and seasons



**Seasonal model**  
SARIMA or Prophet before deep nets

Text, images, audio



**Neural network**  
or a pretrained model + fine-tune

Tiny dataset, must stay readable



**Decision tree**  
print it, show the stakeholder

Any of the above, day one

42

**A dumb baseline**  
mean, majority class, last value

## WORTH MEMORIZING

### 01 Trap: skipping the baseline

A model is only good compared to something.

**predict the mean** Your regression must beat it. Sometimes it barely does.

**majority class** 95% accuracy on 95/5 data is worse than doing nothing.

**last value** Brutally hard to beat on most time series.

→ If the fancy model beats the dumb baseline by 1%, ship the baseline.

### 02 Boosting or neural net?

The question that actually splits modern projects.

#### GRADIENT BOOSTING WHEN

- Data is tabular: rows and columns
- Hundreds to millions of rows
- Features already mean something
- You want feature importance

#### NEURAL NETS WHEN

- Text, images, audio, video
- Raw signal, features unclear
- A pretrained model exists
- You can pay the training bill