How Do We Know, If a Model Is Good?

Quacky has built some clever models.
He has predicted swimming speeds and classified ducks like a pro.
But one day, the wise owl lands beside him and asks:

"How do you know if your model is truly   intelligent…
or just pretending?"

Quacky blinks.
He never thought about that before.

A model might behave perfectly when it sees ducks it already knows…
but when a brand-new duck arrives?

It may get completely confused

This is why evaluation is important, to check whether a model can generalize to real-world ducks it has never met.

Why Performance Must Be Measured

Quacky understands:

"Just like a duck, practicing swimming is not the same as winning a real race!"

A model must:

- Learn real patterns, not just remember
- Handle unseen data confidently
- Avoid silly mistakes

Evaluation is the final exam after all the practice.
If a model only memorizes the training ducks…
- It feels powerful in practice
- But panics in real situations

The owl shakes his head:

"Memorizing is not understanding."

Evaluating Classification Models

Deciding Duck Categories Correctly

When Quacky predicts labels like:
- Winner or Loser
- Fast or Slow
- Duck or Not-a-duck

He must check how often he is right.

The owl shows him a magical scoreboard:

The Confusion Matrix

It compares what is predicted vs what is actually true.

Each box represents a different type of decision:

- **True Positive (TP):** Correctly identifies a real duck

- **True Negative (TN):** Correctly rejects something that is not a duck

- **False Positive (FP):** Mistakes a floating rock for a duck

- **False Negative (FN):** Fails to recognize a real duck in disguise

Quacky thinks:

"I want lots of TP and TN...
and as few wrong guesses as possible!"

Accuracy - Quacky's First Score

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

Higher accuracy means the model makes better decisions.
Simple. Clear. Rewarding.
But the owl whispers

"Someday you will meet more advanced scores
— Precision and Recall — for trickier situations."

Evaluating Regression Models

Seeing how close the predictions are to the truth

When Quacky predicts numbers like:

- Swimming speed
- Flying distance
- Weight next week

Predictions will not always be exact - ducks are unpredictable creatures!

So Quacky measures how big the mistakes are using formulas.

- y_i = Actual value

- $\hat{y}_i$ = Predicted value

- n = Number of observations

The owl introduces two useful evaluation tools:

MAE = Mean Absolute Error
Counts every mistake equally
A smaller value means more accurate

$$MAE = \frac{1}{n} \Sigma \left| y - \hat{y} \right|$$

Divide by the total number of data points

Predicted output value

Actual output value

Sum of

The absolute value of the residual

$$\text{MSE} = \boxed{\frac{1}{n} \sum_{i=1}^{n}} \boxed{(Y_i - \hat{Y_i})}^{\boxed{2}}$$

Mean **Error** **Squared**

MSE = Mean Squared Error
Punishes big mistakes extra strongly
Even smarter about catching big prediction errors

The Danger of Overfitting

When a model becomes too perfect… and too useless

This is a serious warning:

If a model:

- Scores amazingly on the training ducks
- But performs terribly on the test ducks

Then it has overfitted.
It is memorized instead of learning.
Like a duck that memorizes every practice question, but panics in the real exam because the questions look different
Overfitting = Overconfidence without understanding
Quacky fluffs his feathers:
"A model must be prepared for all kinds of ducks!"

What Quacky Understands Now

- Evaluation checks intelligence on new, unseen ducks
- Classification models are judged by how many correct decisions they make
- Regression models are judged by how small their errors are
- Overfitting means the model practiced too much, learned too little

Quacky proudly concludes:
        "A model should be trained well, but judged fairly on what it has never seen before!"

The owl smiles and flies away, leaving Quacky ready not just to build models, but to believe in the right ones.