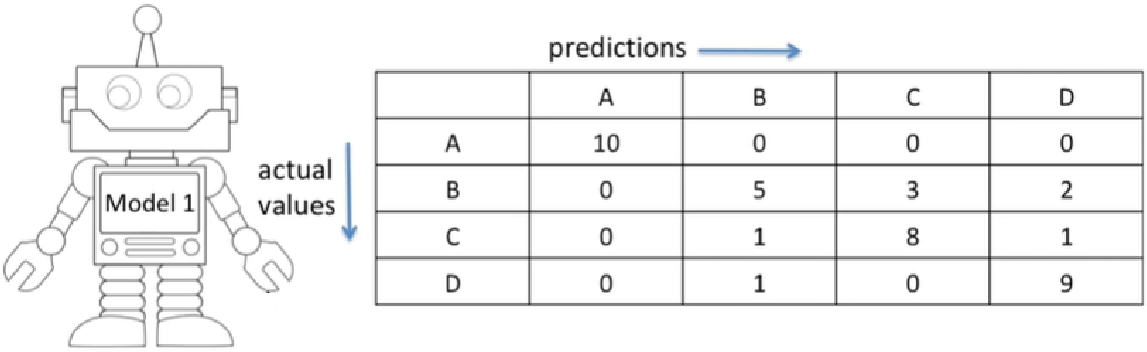
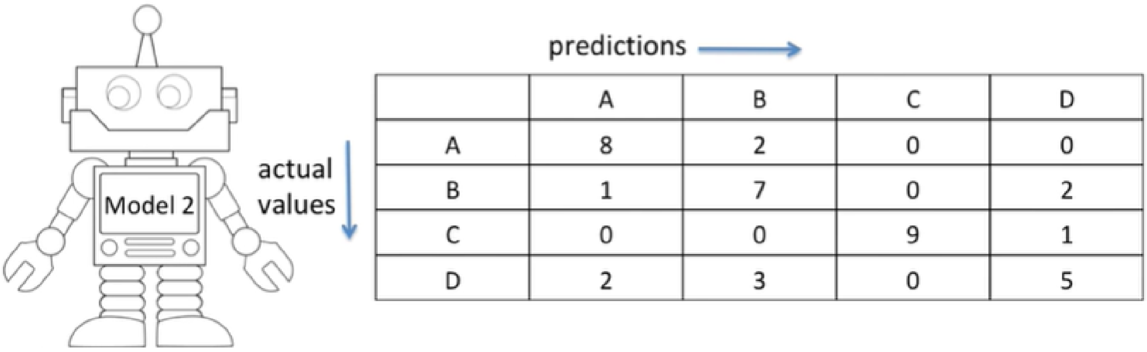
다중분류모델 성능측정

(1) Confusion Matrix

Which model performs better?


Model 1
predictions
actual values

	A	B	C	D
A	10	0	0	0
B	0	5	3	2
C	0	1	8	1
D	0	1	0	9

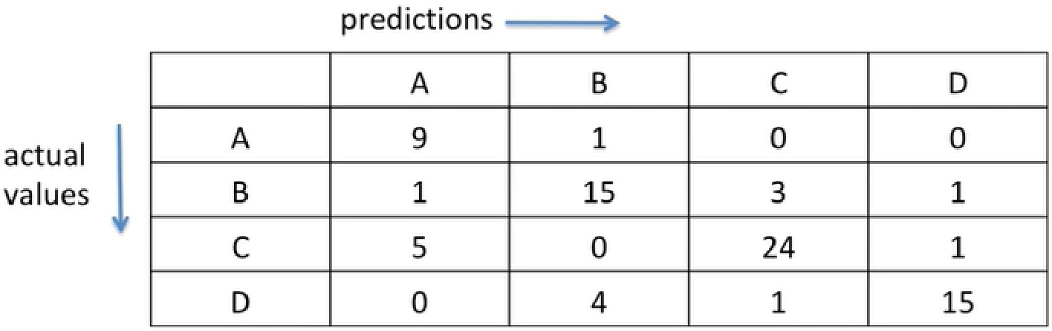

Model 2
predictions
actual values

	A	B	C	D
A	8	2	0	0
B	1	7	0	2
C	0	0	9	1
D	2	3	0	5

(2) Performance Measures

- Accuracy
- Precision
- Recall
- F1 score

- TP (True Positive)
- TN (True Negative)
- FP (False Positive)
- FN (False Negative)

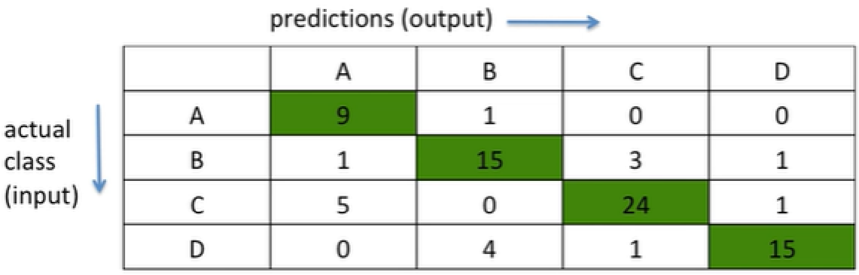


predictions

actual values

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

1) True Positive



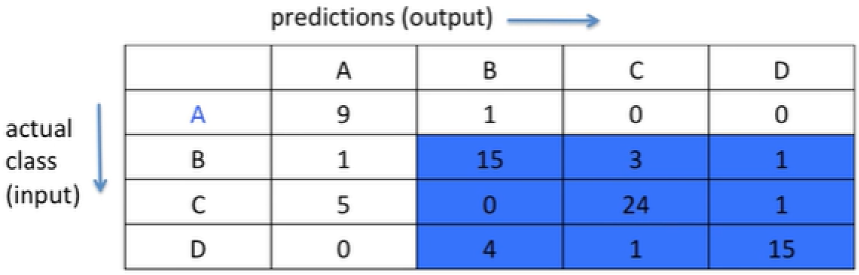
predictions (output)

actual class (input)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

correctly identified prediction for each class

2) True Negative for A



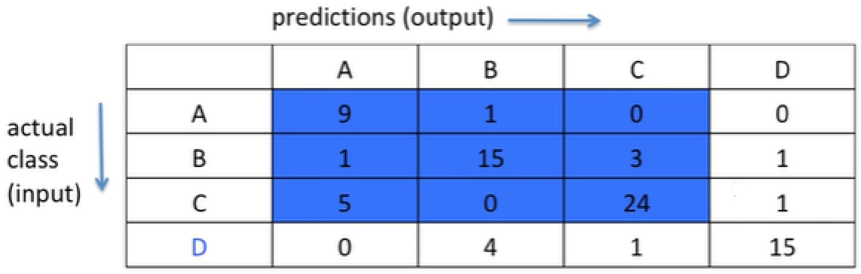
predictions (output)

actual class (input)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

correctly rejected prediction for certain class (A)

3) True Negative for D



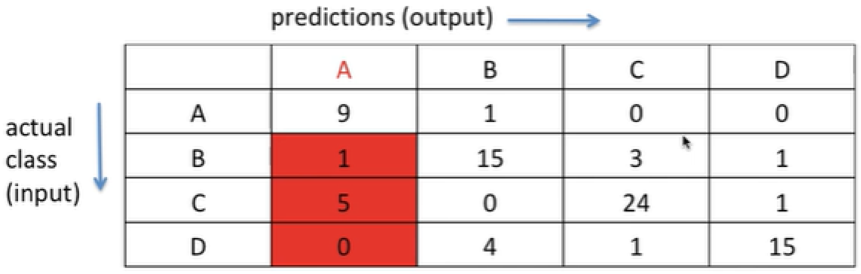
predictions (output) →

actual class (input) ↓

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

correctly rejected prediction for certain class (D)

4) False Positive for A



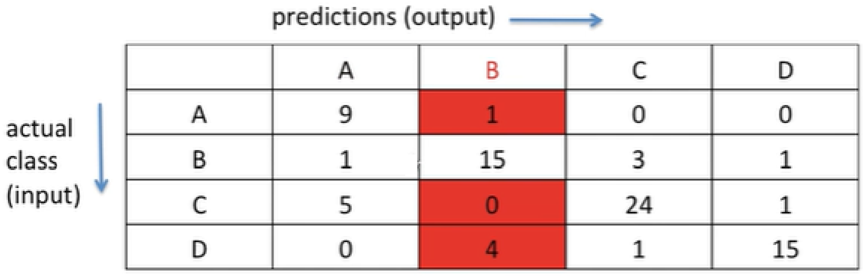
predictions (output) →

actual class (input) ↓

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

incorrectly identified predictions for certain class (A)

5) False Positive for B



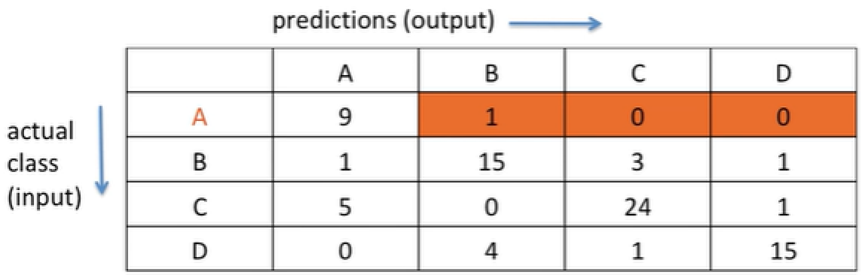
predictions (output) →

actual class (input) ↓

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

incorrectly identified predictions for certain class (B)

6) False Negative for A



predictions (output)

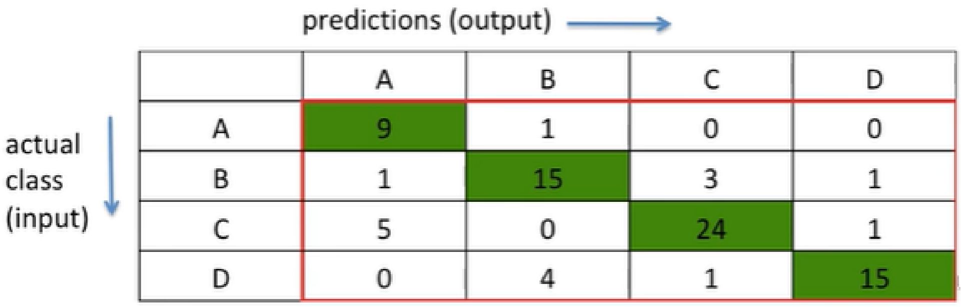
actual class (input)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

incorrectly rejected for certain class (A)

(3) Accuracy

- Accuracy is calculated as the total number of correct predictions divided by the total number of dataset



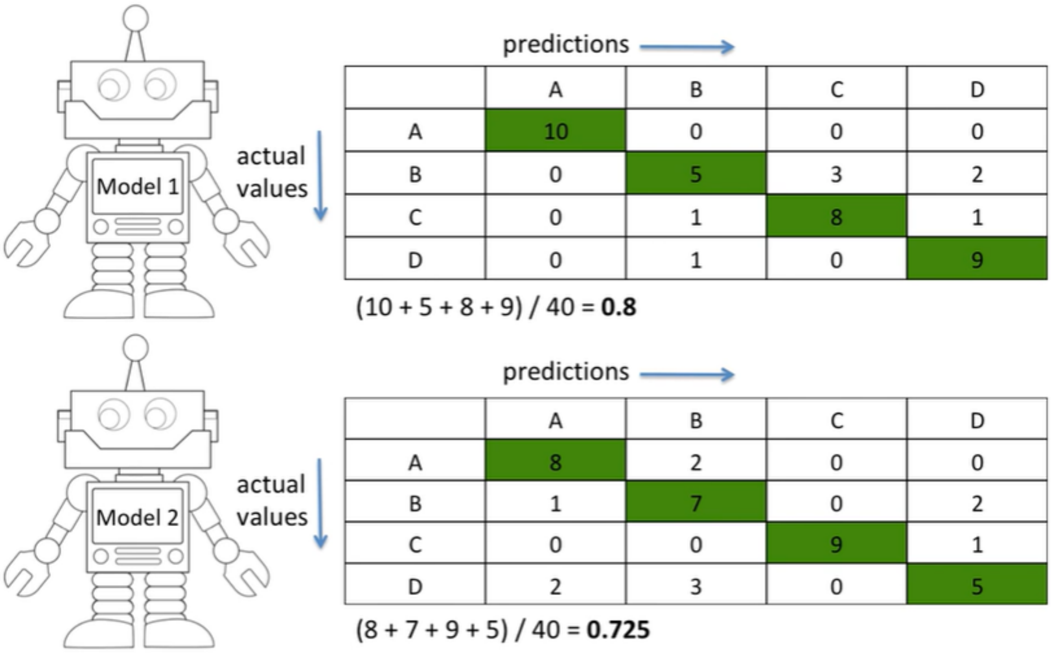
predictions (output)

actual class (input)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

correctly identified prediction for each class / total dataset
9 + 15 + 24 + 15 / 80
accuracy = 0.78

Accuracy Comparison


Model 1
actual values
predictions
Model 2
actual values
predictions

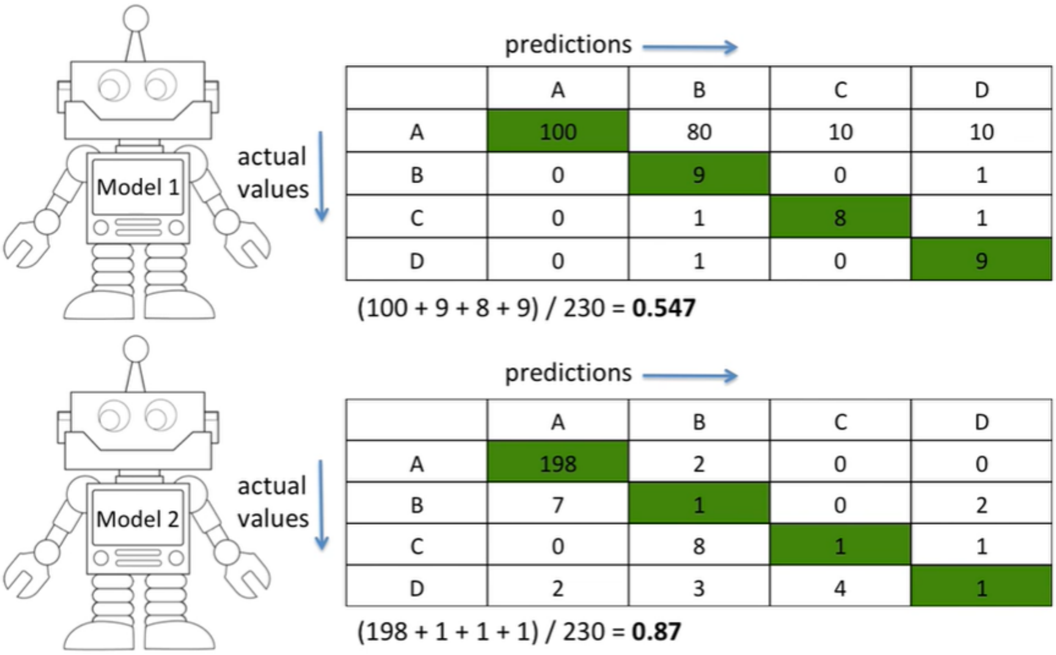
	A	B	C	D
A	10	0	0	0
B	0	5	3	2
C	0	1	8	1
D	0	1	0	9

(10 + 5 + 8 + 9) / 40 = **0.8**

	A	B	C	D
A	8	2	0	0
B	1	7	0	2
C	0	0	9	1
D	2	3	0	5

(8 + 7 + 9 + 5) / 40 = **0.725**

Accuracy on imbalanced data misleads performance


Model 1
actual values
predictions
Model 2
actual values
predictions

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

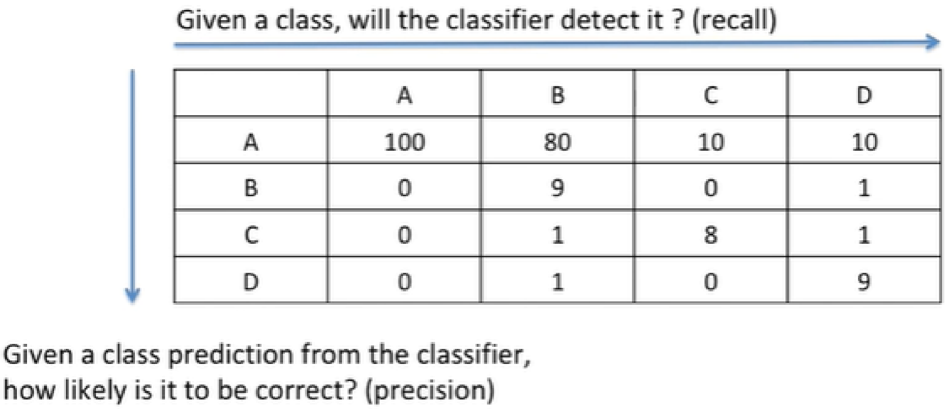
(100 + 9 + 8 + 9) / 230 = **0.547**

	A	B	C	D
A	198	2	0	0
B	7	1	0	2
C	0	8	1	1
D	2	3	4	1

(198 + 1 + 1 + 1) / 230 = **0.87**

(4) F1 Score

F1 score is good metric when data is imbalanced



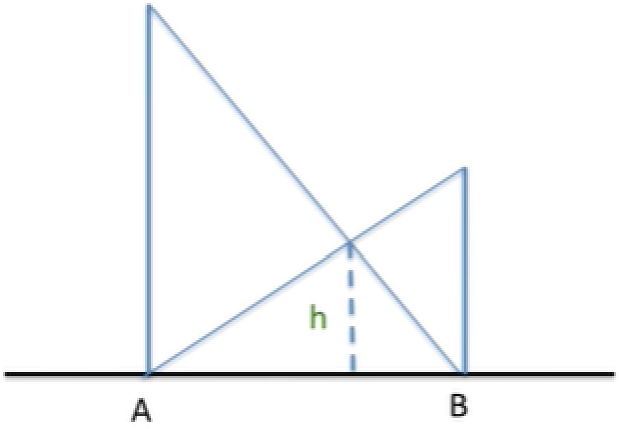
Given a class, will the classifier detect it ? (recall)

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

Given a class prediction from the classifier,
how likely is it to be correct? (precision)

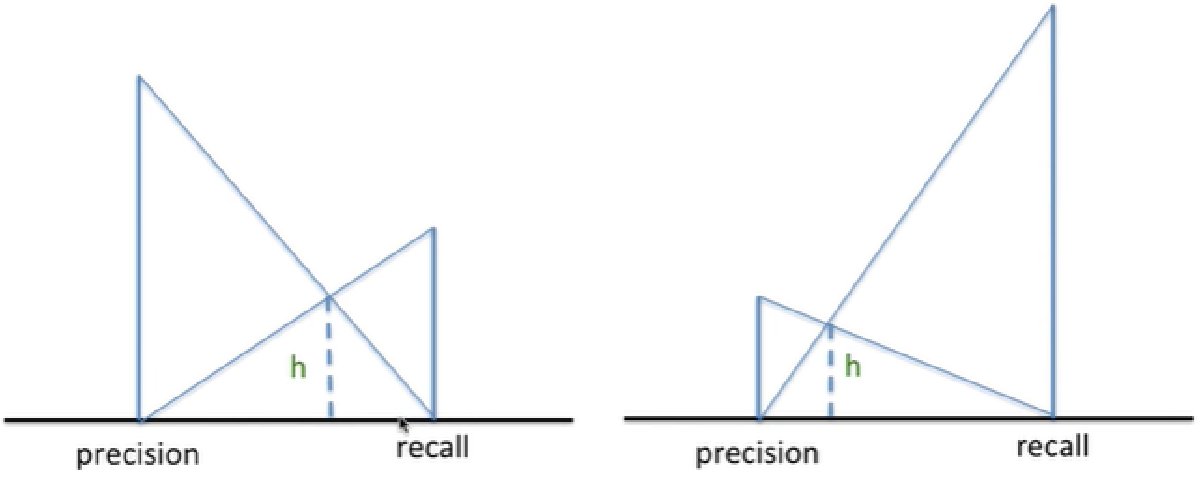
F1 Score is harmonic mean of recall and precision

※ Harmonic Mean



h is half the harmonic mean

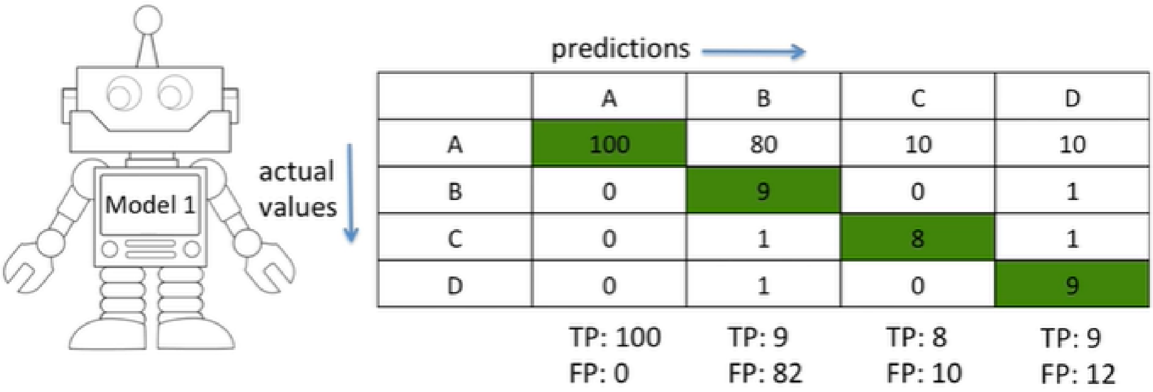
Harmonic Mean punishes extreme value more



h is half the harmonic mean

F1 Score = 2 x $\frac{Precision*Recall}{Precision+Recall}$

Precision of Model1 (macro average)

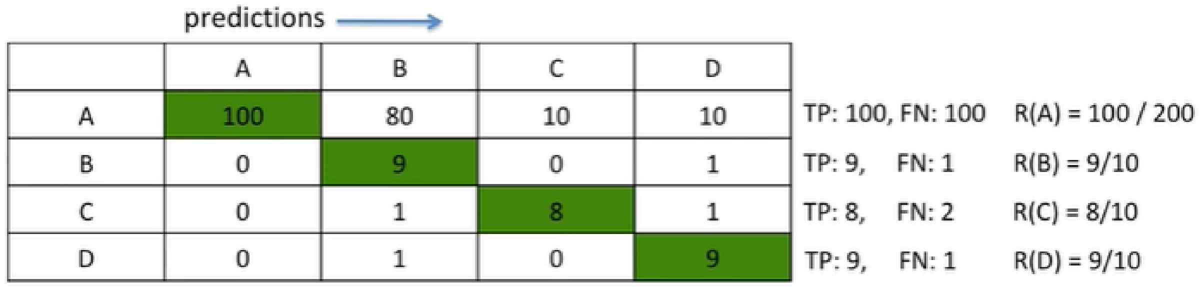

Model 1
actual values
predictions

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9
	TP: 100 FP: 0	TP: 9 FP: 82	TP: 8 FP: 10	TP: 9 FP: 12

Precision = TP / (TP + FP)   P(A) =1   P(B) = 9/91   P(C) = 8/18   P(D) = 9 / 21

average precision = P(A) + P(B) + P(C) + P(D) / 4 = 0.492

Recall of Model1 (macro average)



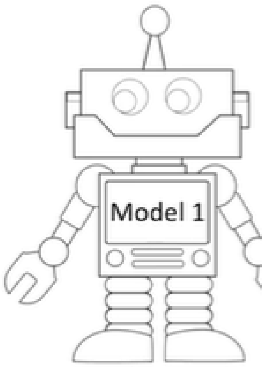
predictions →

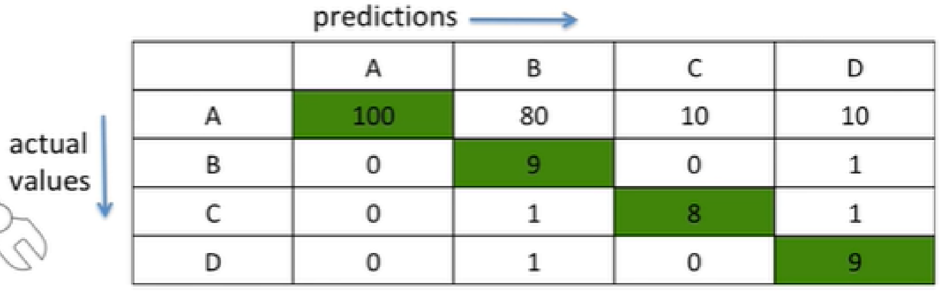
	A	B	C	D		
A	100	80	10	10	TP: 100, FN: 100	R(A) = 100 / 200
B	0	9	0	1	TP: 9, FN: 1	R(B) = 9/10
C	0	1	8	1	TP: 8, FN: 2	R(C) = 8/10
D	0	1	0	9	TP: 9, FN: 1	R(D) = 9/10

Recall = TP / (TP + FN)

average recall = R(A) + R(B) + R(C) + R(D) / 4 = 0.775

F1 Score of Model1





predictions →

actual values ↓

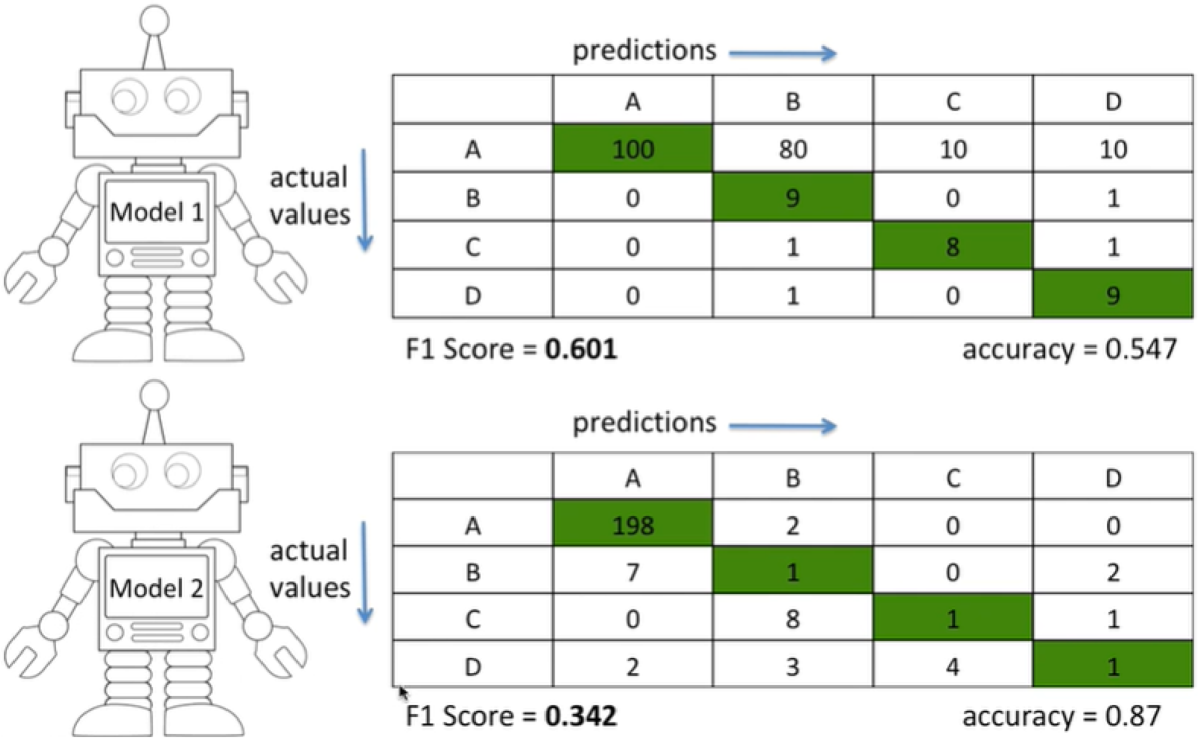
	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

$$\text{F1 Score} = 2 \text{ X } \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

$$2 \text{ X } \frac{0.492 \text{ X } 0.775}{0.492 + 0.775}$$

0.601

F1 Score on imbalanced data


Model 1
predictions
actual values

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

F1 Score = **0.601** accuracy = 0.547

Model 2
predictions
actual values

	A	B	C	D
A	198	2	0	0
B	7	1	0	2
C	0	8	1	1
D	2	3	4	1

F1 Score = **0.342** accuracy = 0.87

Model1 predicts well on multiple class classification on imbalanced given data,

and F1 score is the metric to quantify its performance.