SH Dr. Soufiane Hourri

drsoufiane.com

DUT · SEMESTER 3 · COMPUTER SCIENCE

Machine Learning

Session 04 · Logistic regression: predicting a class

Dr. Soufiane Hourri, PhD, HDR

Associate Professor & AI Consultant

École Supérieure de Technologie de Safi · Cadi Ayyad University

2026-2027

Today's session

01 From counts to probabilities

02 Logistic regression

03 Evaluating a classifier

04 Choosing the threshold

05 Lab: predicting the sex of a penguin

TODAY'S FOCUS

A classifier gives a probability, then a decision. We learn to compute the first, choose the second, and measure both.

Learning objectives

By the end of this session, you will be able to

1. Compute probabilities, conditional probabilities and odds from a table
2. Explain how the sigmoid turns a score into a probability
3. Build a confusion matrix; compute accuracy, precision, recall, F1
4. Compare a classifier with a baseline
5. Choose a decision threshold for a given use

PREREQUISITES

- Session 02: split, encoding, pipelines
- Session 03: linear model, coefficients
- A calculator

Warm-up from Session 03

QUESTION

A linear regression has $R^2 = 0.85$ on its training data. What must we check before trusting it?

A

Nothing: R^2 cannot be negative

B

That 85% of the predictions are exact

C

Its score on the test set, compared with a baseline

D

That the test set was used to fit it

01

CHAPTER

From counts to probabilities

Probabilities, conditional probabilities and odds, computed from a table of counts.

Regression or classification?

Regression

TARGET

A number: price, mass, consumption

EVALUATION

MAE, RMSE, R^2

MODEL

LinearRegression

Binary classification

TARGET

One of two classes, coded 1 and 0

EVALUATION

Confusion matrix, accuracy, precision, recall

MODEL

LogisticRegression

Today's examples

TARGET

Titanic: survived (1) or not (0)

EVALUATION

Penguins: male (1) or female (0)

MODEL

Same scikit-learn workflow

A probability is a proportion

DEFINITION

The probability of an event is the proportion of cases in which it happens: P(event) = count / total. It is always between 0 and 1.

EXAMPLE

342 of 891 Titanic passengers survived:
P(survived) = 342 / 891 = 0.384.

P(did not survive) = 1 − 0.384 = 0.616.

TWO RULES

$0 \leq P(A) \leq 1$

$P(\text{not } A) = 1 - P(A)$

A contingency table: sex and survival

	Did not survive (0)	Survived (1)	Total
Female	81	233	314
Male	468	109	577
Total	**549**	**342**	**891**

Table 1. Titanic passengers (n = 891). Data: seaborn "titanic" dataset. pd.crosstab(titanic["sex"], titanic["survived"]).

Conditional probability

PROBLEM

Using Table 1, compute the probability of surviving for a woman and for a man. Notation: P(survived | female), read "given female".

REASONING

1. Restrict to the group, then divide by the size of the group
2. Women: 233 survivors among 314 → 233 / 314
3. Men: 109 survivors among 577 → 109 / 577
4. Overall: 342 / 891 = 0.384, between the two

RESULT

P(survived | female) = 0.742

P(survived | male) = 0.189

Sex is very informative

Odds and log-odds

$$odds = p \,/\, (1 - p) \qquad \text{log-odds} = \ln(p \,/\, (1 - p))$$

SYMBOLS

p : probability of the event, between 0 and 1

odds: from 0 to +∞; odds = 1 when p = 0.5

log-odds: from −∞ to +∞; 0 when p = 0.5

TITANIC

Women: 233 / 81 = 2.88, ln(2.88) = 1.06

Men: 109 / 468 = 0.233, ln(0.233) = −1.46

Knowledge check

QUESTION

216 passengers travelled in first class; 136 of them survived. What are their odds of surviving?

A

0.63

B

1.70

C

0.59

D

136 / 891 = 0.15

02

CHAPTER

Logistic regression

A linear score, squeezed into a probability by the sigmoid.

Why not linear regression on 0 and 1?

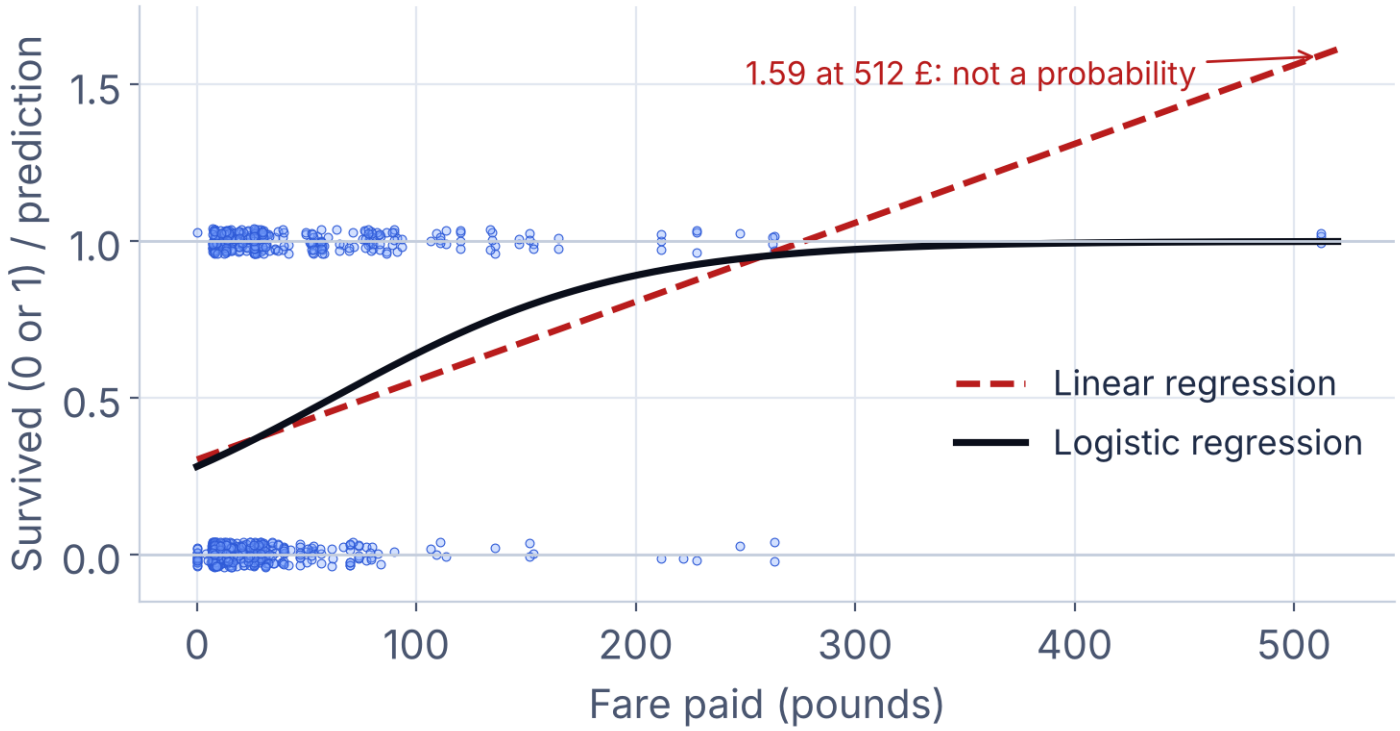


Figure 1. Titanic passengers (n = 891), survived against fare. Fitted with scikit-learn 1.9.1.

WHAT TO NOTICE

- The line predicts 1.59 at 512 £: impossible as a probability
- The logistic curve stays between 0 and 1
- It flattens near 0 and near 1

The logistic regression model

$$z = b_0 + b_1 x_1 + \ldots + b_p x_p \qquad p = 1 / (1 + e^{-z})$$

SYMBOLS

z : linear score = log-odds

p : probability of class 1

$e \approx 2.718$; key e^x on a calculator

INTUITION

z = 0 → p = 0.5; z large → p near 1; z very negative → p near 0.

Same linear formula as Session 03, read through the sigmoid.

The sigmoid function

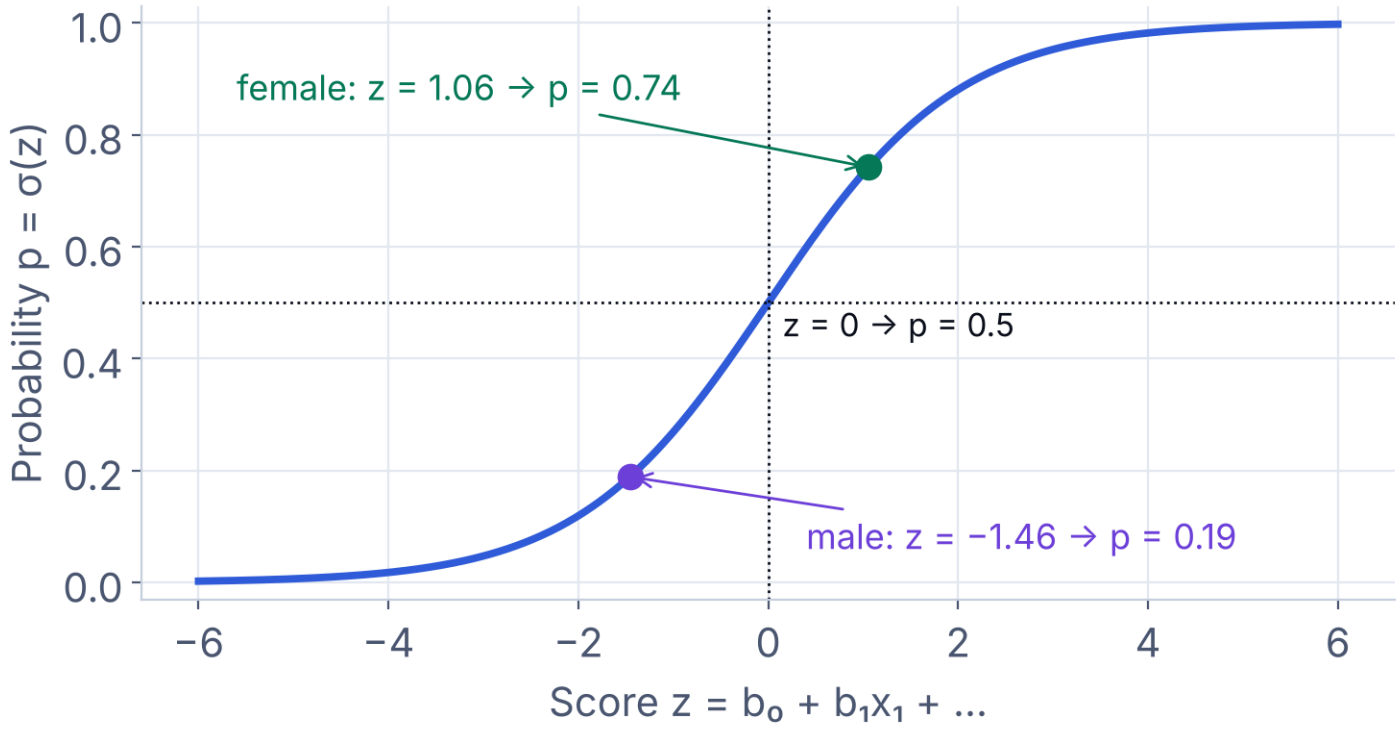


Figure 2. $\sigma(z) = 1 / (1 + e^{-z})$. Points: log-odds of survival for Titanic men and women (Table 1).

WHAT TO NOTICE

- Output always between 0 and 1
- z = 0 gives exactly 0.5
- Men: z = −1.46 → 0.19
- Women: z = 1.06 → 0.74

A logistic model computed by hand

PROBLEM

Model: $z = b_0 + b_1 \times \text{female}$ (female = 1 for a woman, 0 for a man). Use Table 1 to find b_0 and b_1.

REASONING

1. Man: $z = b_0$ = log-odds of men = $\ln(109 / 468) = \ln(0.233)$
2. Woman: $z = b_0 + b_1 = \ln(233 / 81) = \ln(2.88)$
3. $b_1 = 1.06 - (-1.46)$
4. Check: $\sigma(-1.46) = 0.19$, $\sigma(1.06) = 0.74$

RESULT

$b_0 = -1.46$

$b_1 = 2.51$

$z = -1.46 + 2.51 \times \text{female}$

Reading a coefficient: the odds ratio

DEFINITION

e^b is the odds ratio: the odds are multiplied by e^b when the feature increases by 1, all else being equal. b > 0 raises the probability, b < 0 lowers it.

EXAMPLE

b_1 = 2.51 → e^2.51 = 12.4. A woman's odds of survival were about 12 times a man's (2.88 / 0.233 = 12.4).

CAREFUL

Odds, not probabilities: the probability went from 0.19 to 0.74, about ×3.9, not ×12.

LogisticRegression in scikit-learn

CODE · PYTHON

```python
from sklearn.linear_model import (
    LogisticRegression)

titanic = sns.load_dataset("titanic")
X = (titanic[["sex"]] == "female").astype(int)
y = titanic["survived"]
lr = LogisticRegression().fit(X, y)
print(lr.intercept_, lr.coef_)

new = pd.DataFrame({"sex": [0, 1]})
print(lr.predict_proba(new)[:, 1])
```

OUTPUT

```
[-1.43] [[2.45]]

[0.193 0.734]
```

Rounded. Man first, then woman. Close to the hand calculation (small default penalty).

From probability to decision: the threshold

- The model outputs a probability p for class 1
- Decision: predict 1 if p ≥ threshold, else 0
  - Default threshold in scikit-learn's predict: 0.5
- Sex-only model: women (p = 0.73) → survived; men (p = 0.19) → did not survive
- The threshold is a choice: Chapter 4

KEY TAKEAWAY

predict_proba gives probabilities; predict applies the threshold 0.5.

Knowledge check

QUESTION

A model gives z = −1 + 0.5 × x. For x = 4, what is the probability of class 1?

A

About 0.73, so class 1

B

1, so class 1

C

About 0.27, so class 0

D

Exactly 0.5

Break · 15 minutes

After the break: how do we know if a classifier is good?

03

CHAPTER

Evaluating a classifier

Counting the right and wrong answers, by kind of error.

A full Titanic model

PYTHON · SCIKIT-LEARN

```python
train, test = train_test_split(titanic, test_size=0.2,
    random_state=42, stratify=titanic["survived"])
num = ["age", "fare", "sibsp", "parch"]
cat = ["sex", "pclass", "embarked"]
prep = ColumnTransformer([
  ("num", make_pipeline(SimpleImputer(strategy="median"),
                        StandardScaler()), num),
  ("cat", make_pipeline(SimpleImputer(strategy="most_frequent"),
                        OneHotEncoder(handle_unknown="ignore")), cat)])
model = make_pipeline(prep, LogisticRegression(max_iter=1000))
model.fit(train[num + cat], train["survived"])
pred = model.predict(test[num + cat])
```

LINES 1-2

Stratified split: 712 train, 179 test

LINES 3-9

Session 02 preparation

LINES 10-12

Same pipeline, new final model

The confusion matrix

	Predicted 0 (did not survive)	Predicted 1 (survived)	Total
Actual 0 (did not survive)	**TN = 98**	**FP = 12**	110
Actual 1 (survived)	**FN = 23**	**TP = 46**	69
Total	121	58	179

Table 2. Titanic test set (179 passengers). TP true positive, TN true negative, FP false positive, FN false negative.

Four metrics from the confusion matrix

$$\text{accuracy} = (\text{TP} + \text{TN}) / \text{total} \quad \text{precision} = \text{TP} / (\text{TP} + \text{FP}) \quad \text{recall} = \text{TP} / (\text{TP} + \text{FN})$$
$$\text{F1} = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$$

MEANING

Accuracy: share of correct answers

Precision: when it says 1, how often right?

Recall: of the real 1s, how many found?

F1: balance of precision and recall

INTUITION

Each metric is a proportion between 0 and 1.

Precision is hurt by false alarms (FP); recall by misses (FN).

Computing the four metrics

PROBLEM

From Table 2 (TN 98, FP 12, FN 23, TP 46), compute accuracy, precision, recall and F1.

REASONING

1. Accuracy = (46 + 98) / 179 = 144 / 179
2. Precision = 46 / (46 + 12) = 46 / 58
3. Recall = 46 / (46 + 23) = 46 / 69
4. F1 = 2 × 0.793 × 0.667 / (0.793 + 0.667)

RESULT

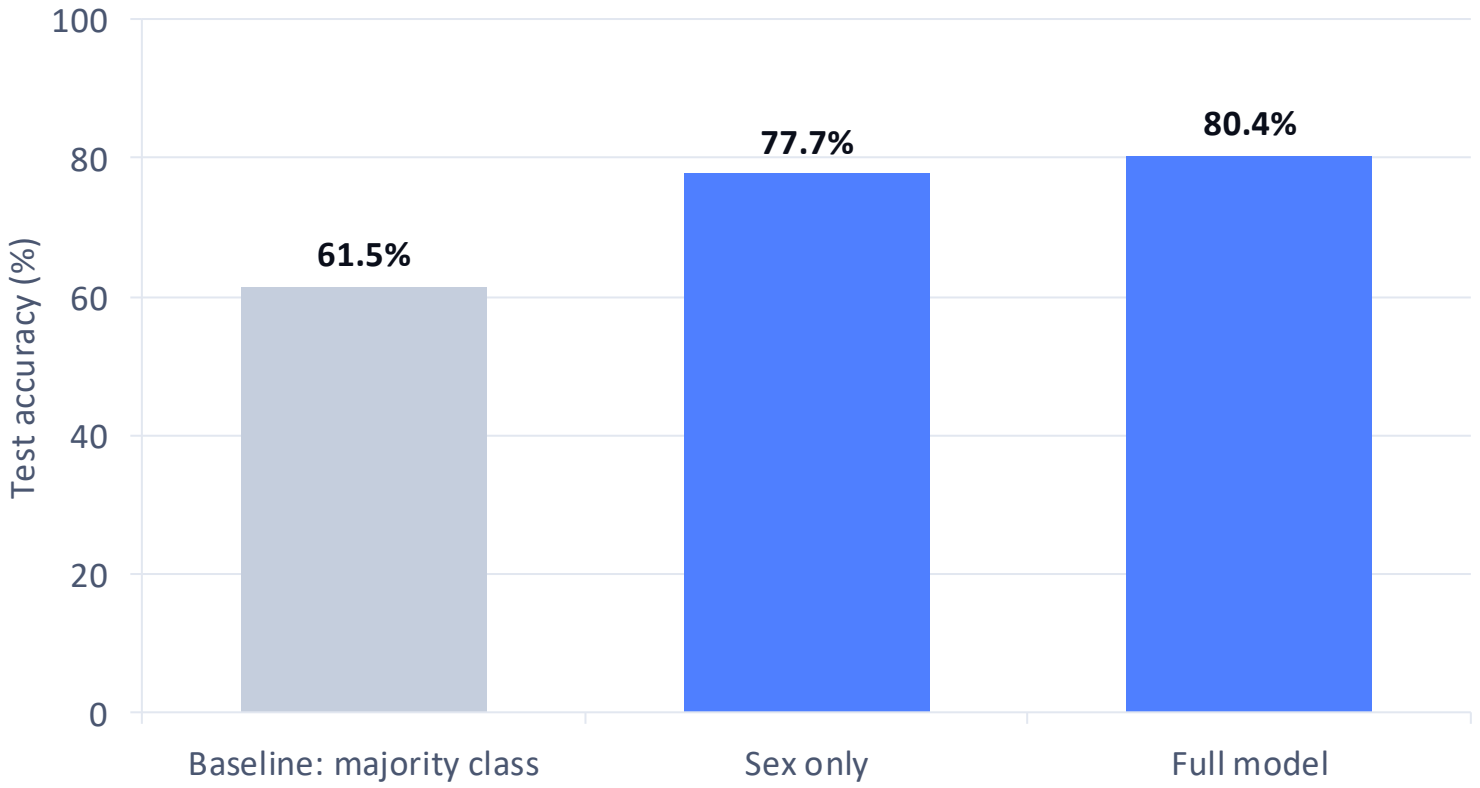
Accuracy 0.804

Precision 0.793

Recall 0.667

F1 0.724

Compare with a baseline



Titanic test set (179 passengers). Baseline: DummyClassifier(strategy="most_frequent"). scikit-learn 1.9.1.

MAIN FINDING

Predicting "did not survive" for everyone is already 61.5% accurate.

The model's real gain is about 19 points.

Trusting accuracy on imbalanced data

✗ INCORRECT APPROACH

"My fraud detector is 99% accurate": on 1,000 transactions with 10 frauds, it answers "not fraud" every time

✓ CORRECTION

Report recall and precision for the rare class, and compare with the majority baseline (here also 99%)

WHY IT FAILS

With 99% of one class, accuracy rewards ignoring the other. Recall = 0 / 10 = 0: not a single fraud is found.

Knowledge check

QUESTION

A disease test: TP = 40, FN = 10, FP = 20, TN = 130. What is the recall?

A

0.67

B

0.85

C

0.80

D

0.33

04

CHAPTER

Choosing the threshold

Precision or recall: the threshold decides, the application chooses.

Precision and recall move in opposite directions

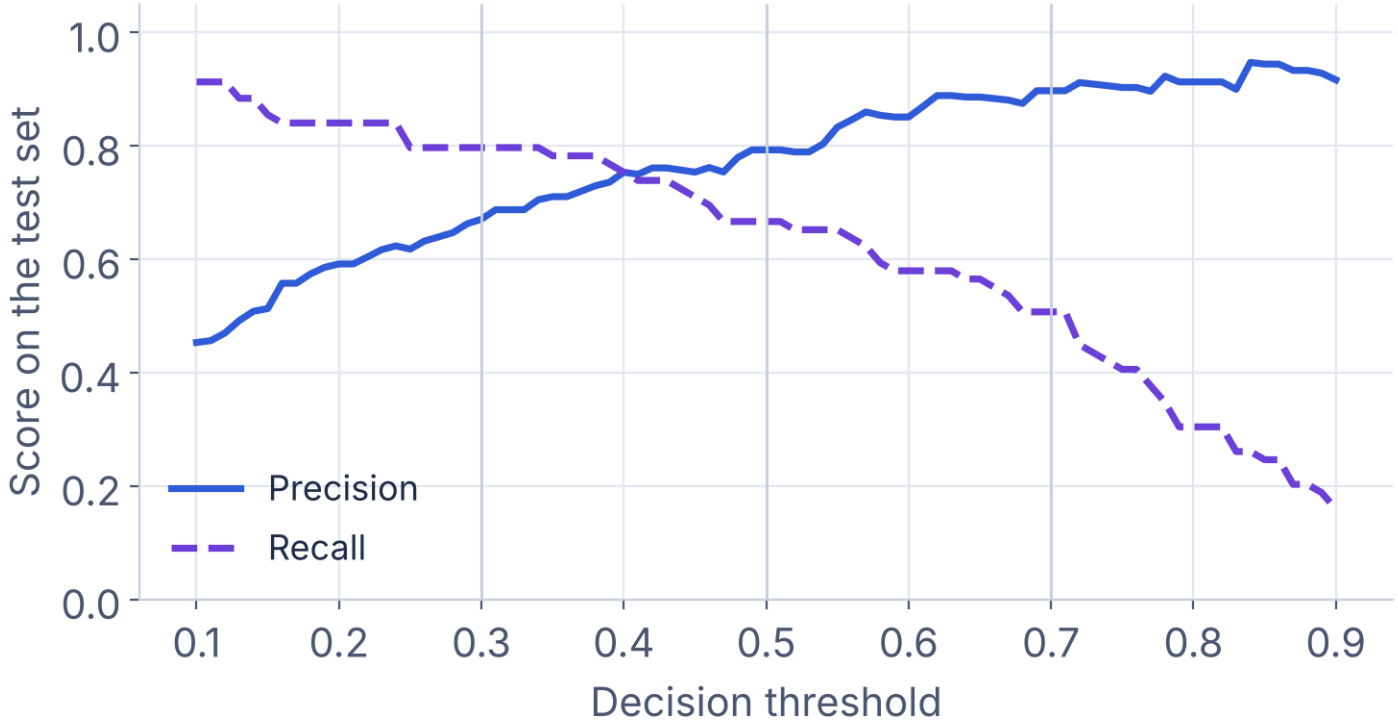


Figure 3. Full Titanic model, test set (179 passengers). Vertical lines: thresholds 0.3, 0.5, 0.7.

WHAT TO NOTICE

- Lower threshold: recall up, precision down
- Higher threshold: precision up, recall down
- No threshold maximises both

Three thresholds, three behaviours

Threshold	TP	FP	FN	TN	Accuracy	Precision	Recall
0.3	55	27	14	83	0.771	0.671	0.797
0.5	46	12	23	98	0.804	0.793	0.667
0.7	35	4	34	106	0.788	0.897	0.507

Table 3. Full Titanic model on the test set (179 passengers), for three decision thresholds.

Which threshold for which use?

Lower threshold (e.g. 0.3)

EFFECT

Favour recall: find most positives

WHEN

Missing a positive is costly

EXAMPLES

Medical screening, fraud alerts

Default 0.5

EFFECT

Balance

WHEN

Errors cost about the same

EXAMPLES

First model, F1 as summary

Higher threshold (e.g. 0.7)

EFFECT

Favour precision: few false alarms

WHEN

A false alarm is costly

EXAMPLES

Spam filter, automatic sanctions

Applying your own threshold

CODE · PYTHON

```python
proba = model.predict_proba(test[num + cat])[:, 1]
pred30 = (proba >= 0.3).astype(int)

y_test = test["survived"]
print(confusion_matrix(y_test, pred30))
print(precision_score(y_test, pred30))
print(recall_score(y_test, pred30))
```

OUTPUT

```
[[83 27]
 [14 55]]
0.671
0.797
```

Rounded. Same as the 0.3 row of Table 3.

05

LAB

Lab: predicting the sex of a penguin

Can body measurements tell a male penguin from a female?

LAB · GUIDED CODE

Lab · Preparing the data

CODE · PYTHON

```python
df = pd.read_csv("penguins.csv")
df = df.dropna(subset=["sex"])
df["male"] = (df["sex"] == "Male").astype(int)
print(len(df), df["male"].mean())

train, test = train_test_split(df, test_size=0.2,
    random_state=42, stratify=df["male"])
print(len(train), len(test))
```

OUTPUT

```
333 0.505
266 67
```

Rounded. Balanced classes: about half are male.

LAB · EXERCISE

Lab · Male or female?

GOAL

Predict male (1) or female (0) with logistic regression; evaluate and interpret.

TIME

35 min

INSTRUCTIONS

1. Baseline accuracy (DummyClassifier)
2. Model with body_mass_g only: accuracy, confusion matrix
3. Interpret the coefficient for 100 g (odds ratio)
4. Why is mass alone weak? (Hint: species)
5. Pipeline: 4 measurements scaled + species one-hot: 4 metrics
6. Thresholds 0.3 and 0.7: precision and recall?

DELIVERABLE

Your notebook, with one sentence per question.

Exit question

QUESTION

A screening test must miss as few sick people as possible. What do you do?

A

Raise the threshold to 0.8

B

Lower the threshold to favour recall

C

Maximise accuracy only

D

Nothing: the threshold does not change the predictions

Recap

KEY TAKEAWAYS

- A probability is a proportion; odds = p / (1 − p)
- Logistic regression: linear score → sigmoid → probability
- e^b is an odds ratio, all else equal
- Confusion matrix → accuracy, precision, recall, F1
- Compare with the baseline; choose the threshold for the use

NEXT

- Session 05: decision trees
- Practice: finish the lab notebook
- Read: ISLP, chapter 4

References

(1) James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning, with Applications in Python, ch. 4 (Classification). Springer. statlearning.com

(2) Diez, D., Çetinkaya-Rundel, M., & Barr, C. D. (2019). OpenIntro Statistics (4th ed.), ch. 3 (Probability) and 9.5 (Logistic regression). openintro.org/book/os

(3) scikit-learn 1.9 User Guide: Logistic regression; Metrics and scoring: classification metrics. scikit-learn.org/stable

(4) Titanic passenger data, as distributed with seaborn (seaborn-data repository, "titanic"). github.com/mwaskom/seaborn-data

(5) Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. doi.org/10.5281/zenodo.3960218

Course page
learn.drsoufiane.com

SH

Thank you · Questions?

Next session: decision trees. Finish the lab notebook for Friday's practice.

Dr. Soufiane Hourri, PhD, HDR

drsoufiane.com