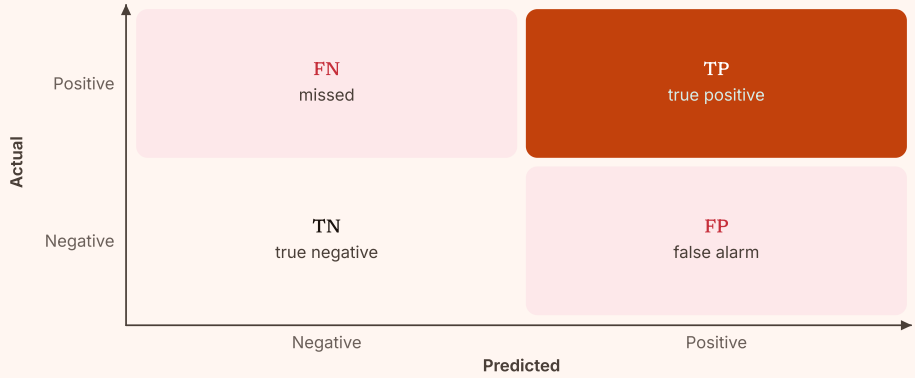
Logistic regression, overfitting and validation

CHEATSHEET

Exam Mon, 7 Dec 2026

EXAM LIKELIHOOD Logistic regression (logit model) **98%** · Confusion matrix, computing accuracy and sensitivity **97%** · Accuracy is misleading with imbalanced classes **95%** · A logit coefficient is a change in log-odds, not in probability **94%** · Odds and odds ratio (e^b) **93%** · Overfitting and generalization **92%**

Confusion matrix



Preventing overfitting with validation

1. Split into train and test
2. Fit on the train set
3. Predict the test set
4. Compare train and test performance

M 01 Logit (formula sheet)

- odds = p/(1 – p); log-odds z = ln(p/(1 – p)) = b0 + b1 ×1 + …
- p = 1/(1 + exp(-z)); odds ratio = exp(b)
- b = change in log-odds, **not** in percentage points
- More units: exp(n b), not n times the odds ratio

M 02 R

- **glm(y ~ x1 + x2, family = binomial, data = d)**
- **predict(fit, newdata, type = "response")** gives probabilities; without type log-odds
- **exp(coef(fit))**, **exp(confint.default(fit))** for odds ratios with a Wald interval (as in the app); **confint(fit)** gives the profile interval, slightly different
- **table(Voorspeld, Waargenomen)** for the confusion matrix
- **anova(m1, m2, test = "Chisq")** for nested models
- ROCR: **prediction()**, **performance()** for ROC curve and AUC

M 03 Reading output

- **z value** and **Pr(>|z|)**: below 0.05 is significant (sign and significance)
- Null deviance (intercept only) and residual deviance (model); df = n – k
- AIC = 2k – 2 ln(L): lower is better
- McFadden pseudo-R^2 = 1 – ln L(model)/ln L(null) (not explained variance)

M 04 Confusion matrix (threshold τ)

	Actual 0	Actual 1
Predicted 0	TN	FN
Predicted 1	FP	TP

Watch the order. R (**table**) puts 0 first: TN top left. The practice exam 7.1 puts 1 first: TP top left, FN top right, FP bottom left, TN bottom right. Read the row and column headings first before you point out TP, FN, FP and TN.

- accuracy = (TP + TN)/n
- sensitivity = TP/(TP + FN); specificity = TN/(TN + FP)
- precision = TP/(TP + FP)
- Lower threshold: sensitivity up, specificity down
- AUC: 0.5 is guessing, 1 is perfect

M 05 Decision rules

- Imbalanced classes: compare accuracy with the baseline (always the majority class); look at sensitivity and precision
- High cost for a missed case: lower threshold
- Overfitting: train much better than test; choose the simpler model
- Judge on a test set or with cross-validation, never on the training data alone
- More flexible model: less bias, more variance

M 06 Question 7.1 (9 p, 200 words)

1. Coefficients: sign, significance, odds ratio, ceteris paribus
2. Calculation with formula and numbers
3. Why accuracy misleads: imbalanced outcome, low sensitivity, consequence for the company (threshold or overfitting only if the case asks for it)

! Pitfalls

- p = 0.06 is not a 6% chance of H0
- "2.75 times the chance" is wrong: it is odds
- 0.5 is a habit, not a rule
- A p-value says nothing about the size of the effect
- MLE properties hold for large n