COURSEWORK ANALYSIS / 2026

Spam, signals, and the cost of a false alarm

A transparent logistic-regression study of a historical email dataset. The central question: how well can measured message patterns identify spam, and how does the decision change when legitimate mail needs extra protection?

4,601	57	39.4%
LABELED EMAILS	NUMERIC SIGNALS	SPAM IN DATASET

Method in brief

01

Audit

Check class balance, missingness, repeated rows, and feature distributions.

02

Fit

Use an 80/20 stratified split. Select preprocessing and L2 logistic regression with five training-only CV folds.

03

Evaluate

Score the untouched 921-email holdout, then test how a higher decision threshold changes errors.

Holdout result

DEFAULT DECISION THRESHOLD / 0.50

94.0%	0.982	91.2%	95.9%
ACCURACY	ROC-AUC	SPAM RECALL	SPECIFICITY

THE ERRORS BEHIND THE SCORE

23 legitimate emails incorrectly flagged as spam

32 spam emails left in the inbox

THRESHOLDS AND SENSITIVITY

One model. Two decisions.

The model estimates spam probability; a threshold turns that score into an action. Raising the cutoff reduces false alarms, but lets more spam through.

What changed on the holdout

STANDARD CUTOFF

0.50	95.9%	91.2%	23	32
THRESHOLD	NON-SPAM KEPT	SPAM CAUGHT	FALSE ALARMS	MISSED SPAM

HIGH-SPECIFICITY CUTOFF

0.934	99.5%	63.4%	3	133
THRESHOLD	NON-SPAM KEPT	SPAM CAUGHT	FALSE ALARMS	MISSED SPAM

The higher cutoff was selected using nested, duplicate-grouped predictions from training data. The figures above are then measured on the separate holdout. The targeted specificity was 99% in training; observed holdout specificity was 99.5%.

Repeated records matter

There were 391 exact duplicate rows beyond their first occurrence. In the primary random split, 105 holdout rows shared a predictor pattern with training data. I checked whether separation changed when those patterns were handled differently.

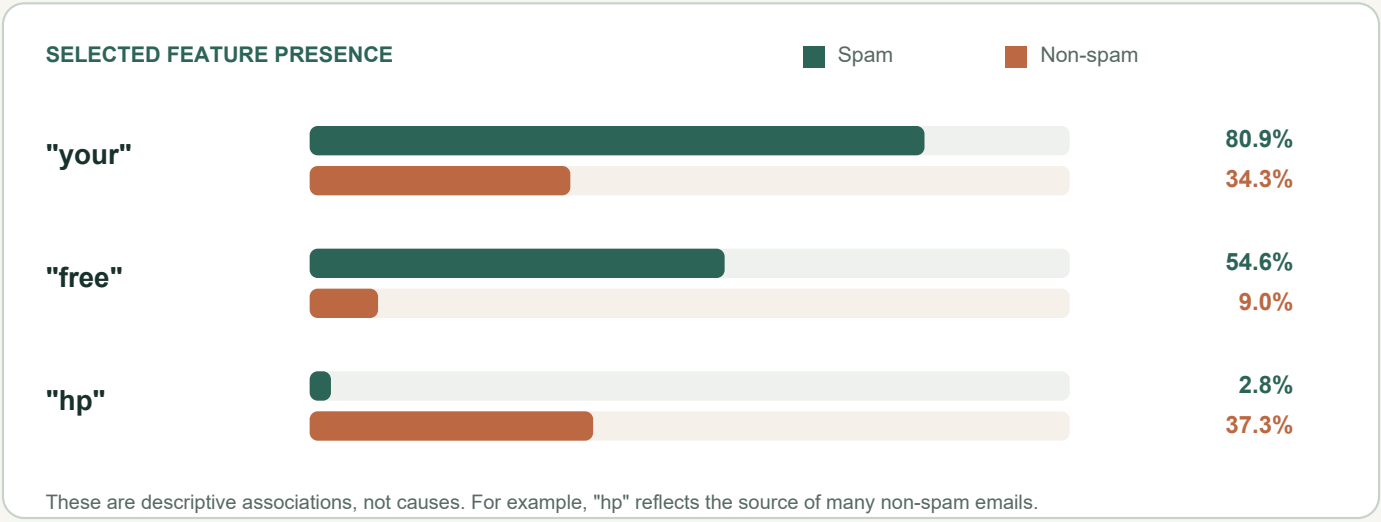
EVALUATION DESIGN	OVERLAP	ACCURACY	ROC-AUC
Primary random holdout	**105**	**94.0%**	**0.982**
Exact rows deduplicated	**2**	**94.4%**	**0.982**
Predictor-group 5-fold CV	**0**	**93.6%**	**0.980**

Designs use different samples or partitions; the comparison is a sensitivity check, not an external validation.

EVIDENCE WITH BOUNDARIES

What the benchmark shows

Some measured patterns differed sharply between the historical spam and non-spam messages. The bars show the share of emails in each class where the named feature was present.



These are descriptive associations, not causes. For example, "hp" reflects the source of many non-spam emails.

Limits that travel with the result

Historical sample

Spambase was donated in 1999. Its language and source organizations do not represent current global email.

No deployment claim

A single held-out partition and sensitivity checks are reproducible benchmark evidence, not proof of future performance in a live mail system.

Data artifacts

Repeated predictor patterns can cross random splits; three repeated patterns also have conflicting labels. Grouped evaluation reduced this overlap to zero.

DATASET & ATTRIBUTION

Hopkins, M., Reeber, E., Forman, G., & Suermondt, J. (1999). Spambase [Dataset]. UCI Machine Learning Repository. doi.org/10.24432/C53G6X. Licensed CC BY 4.0. Analysis and report: Paul Mwangi, 2026.