

Income Level Prediction Using R

TECHNICAL REPORT AND PORTFOLIO DOCUMENTATION

A reproducible comparison of four classifiers and a majority baseline on the historical UCI Adult income classes.

Esad Kopru | Independent learning project | Revised benchmark completed 8 October 2026

86.42%

Decision-tree accuracy on the reused benchmark. The highest observed accuracy among the compared models.

0.9090

Random-forest ROC AUC. The forest also had the highest average precision and lowest aggregate probability error.

16,255

Retained test records, including 3,840 in the positive income class. All metrics are unweighted sample statistics.

The contribution. The revised workflow makes data handling, training-only model selection, fixed thresholds, uncertainty, and output provenance inspectable. Different performance criteria favor different models; the evidence does not establish a universal winner.

Evidence boundary. The official test data were already examined in version 1. Version 2 is a reused historical benchmark, not fresh independent confirmation or a current-income estimator.

Read this report

02-05 Project and methodology
06-11 Results and diagnostics
12 Reproduce and verify
13 Conclusions and limits
14 Sources and evidence map

Sources: project README, version 2 protocol, published results [1-3].

From a learning script to an auditable study

PURPOSE AND PROJECT EVOLUTION

Study question

How do four supervised classifiers compare when data preparation and selection are kept inside training data, while classification, ranking, probability quality, and subgroup errors are reported separately?

Positive class: **>50K**. Negative class: **<=50K**. The threshold and category labels belong to the historical dataset; they do not represent present-day income purchasing power.

Evidence rather than score inflation

Version differences combine several methodological changes. The project does not attribute a score increase to any single correction without a controlled comparison.

What changed

Stage	Documented treatment
Original project	Preserved verbatim. Review identified row-index mismatches, training/evaluation contamination, a log-odds threshold issue, and ROC calculations using class labels.
Version 1	Corrected the workflow. Its source and evidence remain separately archived. Its test inspection limits later claims of independence.
Version 2	Adds grouped selection, revised retention, capital transforms, ridge regularization, a probability forest, conditional uncertainty, and traceable outputs.
Summary release	Adds ten ggplot2 views and eleven derived statistics tables from the same verified aggregates. No model refit or new uncertainty calculation.

Engineering contribution: locked package versions; explicit downloads and input checksums; deterministic configuration; fold-local preprocessing; failure checks; exact numeric serialization; aggregate-only public export; integrity verification.

Sources: code review, preserved snapshots, protocol and validation record [1-4].

Data retention and denominators

OFFICIAL FILES REMAIN SEPARATE

Sample retention and class balance define the denominators

Exclusions: 24 repeated labeled training rows; 26 test rows sharing raw predictor signatures with training



UCI Adult | reused historical benchmark | 16,255 test rows
Retention percentages use each raw partition as denominator. Income percentages use each retained partition; the panels are not additive.
Missing values are handled by fold-fitted preprocessing, not complete-case deletion. Profile matches are not confirmed person identities.

Training retains two rows with conflicting labels for one row signature. Six extra row-signature repetitions remain in the test sample. Signatures are not person identifiers.

Sources: data_audit.csv, cv_metrics.csv, metrics.csv; summary figure 10 [2,5].

Training-only preparation and model selection

FIVE GROUPED FOLDS, ONE DECLARED SELECTION RULE

Data handling

Remove 24 identical complete training records. Exclude 26 test rows matching any original training signature across all 14 raw predictors, without consulting test labels. Retain 32,537 training and 16,255 test records.

Within each analysis fold, learn numeric medians, centering/scaling, category levels, reference categories, and constant-column removal. Map later unseen categories to the learned mode and record the count.

Apply fixed **log1p** transforms to capital gain and loss. Retain numeric education; exclude text education and **fnlwgt**. Use no observation weights. Recorded sex and race remain predictors.

Keep each 14-field predictor signature within one training fold. Approximate class balance across five folds. Every model receives the same encoded predictors.

Model / engine	Declared candidates	Selected
Majority baseline	Constant training prevalence	No tuning
Ridge / glmnet	Lambda: 0.1, 0.01, 0.001, 0.0001	0.0001
Tree / rpart	cp: 0.01, 0.005, 0.001, 0.0005	0.0005
Forest / ranger	mtry: floor(sqrt(p)), floor(p / 3)	floor(p / 3); final 28
Linear SVM / e1071	Cost: 0.1, 1, 10	10

Selection: highest mean fold balanced accuracy; exact ties use declaration order. Refit selected settings on all retained training records. Forest: 500 trees, one thread, minimum split-node size 10.

Decisions: probability >0.5 or oriented SVM margin >0 predicts the positive class. Ties predict the negative class. No threshold optimization or post-test calibration is part of version 2.

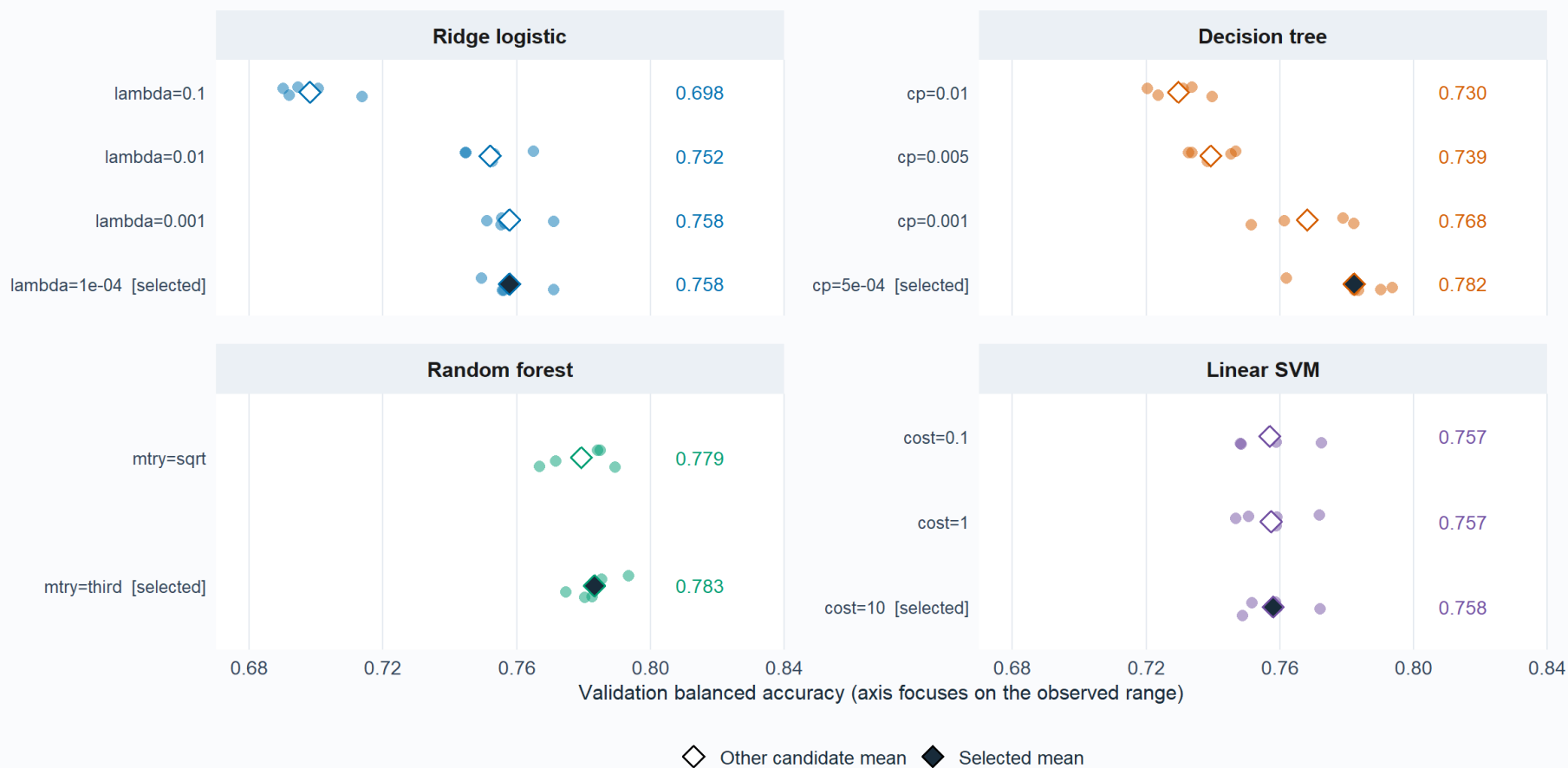
Source: version 2 protocol, selected_parameters.csv and run_config.txt [2-3].

Training cross-validation preferred the forest

SETTINGS SELECTED BEFORE VERSION 2 TEST EVALUATION

All 13 trained configurations in the validation search

Training folds only | Circles: five fold scores; diamonds and right labels: arithmetic means



UCI Adult | reused historical benchmark | 16,255 test rows

This chart uses training cross-validation, not the 16,255-row test set. Filled diamonds and [selected] identify each family's chosen configuration.

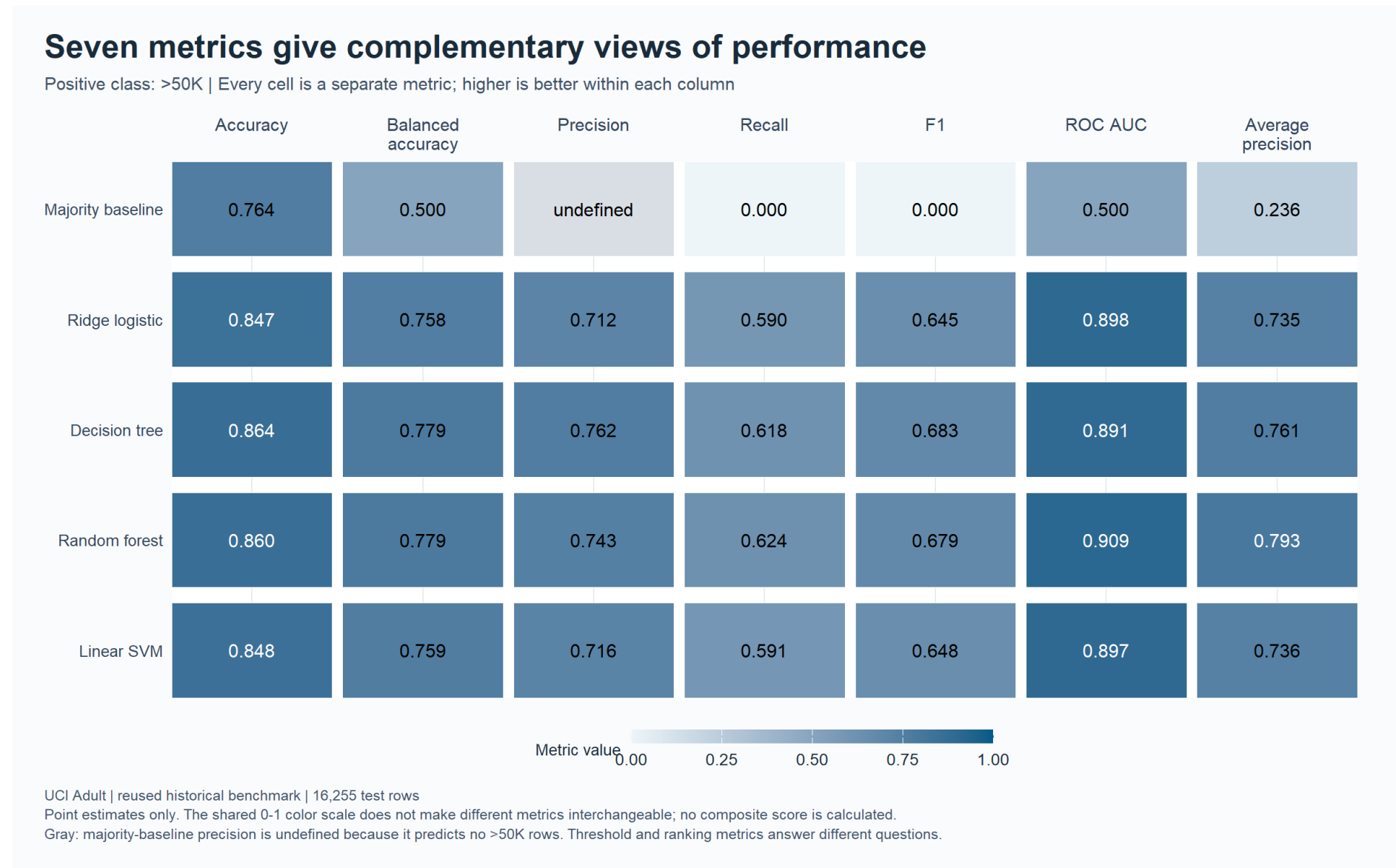
Fold spread is descriptive, not a confidence interval. Selection reuses these folds; tested grid boundaries do not establish a global optimum.

Mean balanced accuracy: forest 78.32%, tree 78.23%; the difference is 0.095 percentage points. The 13 trained configurations plus baseline produce 70 candidate-fold fits.

Sources: cv_metrics.csv, cv_summary.csv; summary figure 07 [2,5].

Classification and ranking favor different models

FIVE MODELS EVALUATED ON THE SAME 16,255 RECORDS

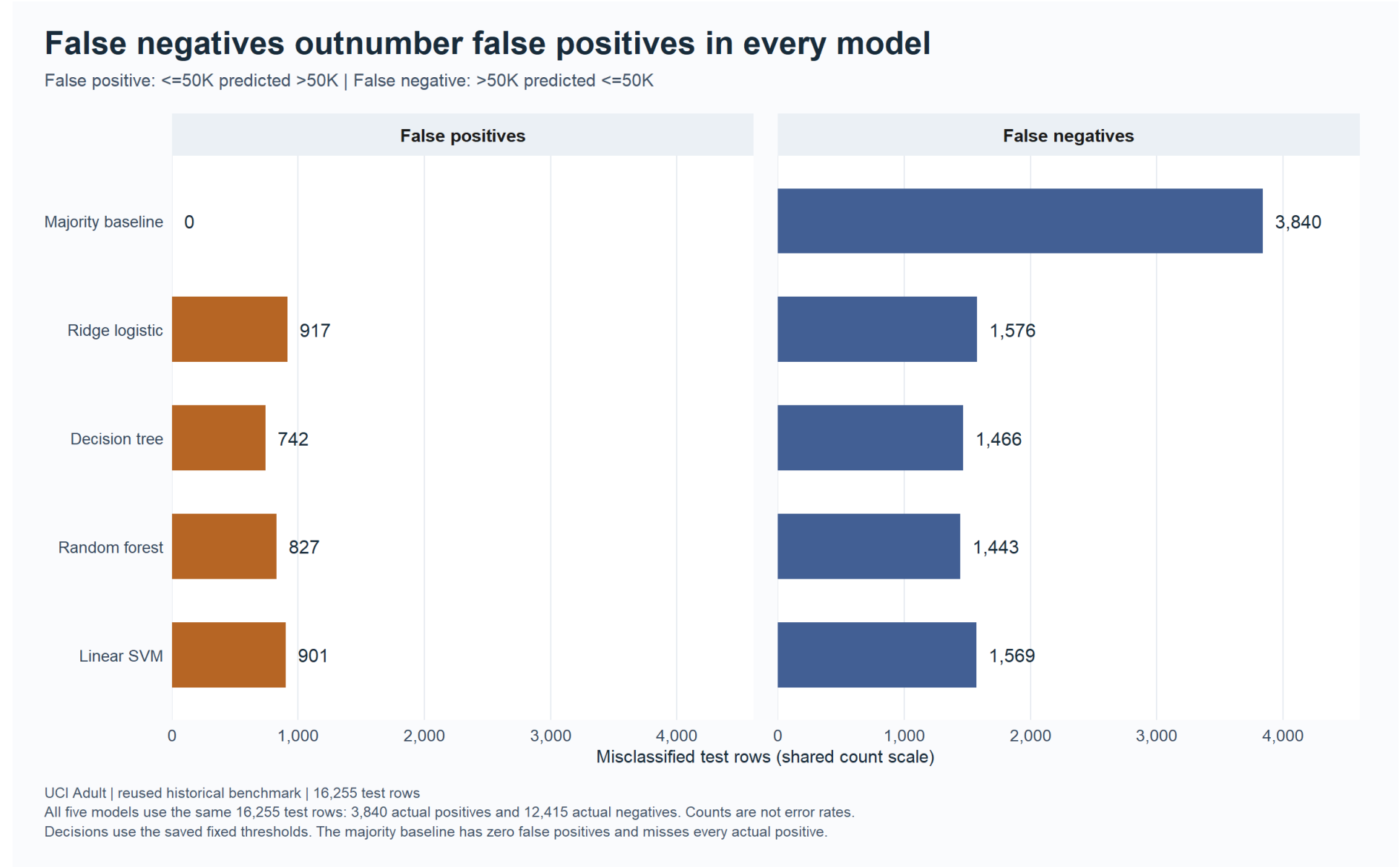


The tree has the highest observed accuracy and balanced accuracy. The forest has the highest ROC AUC and average precision. These point estimates do not establish universal superiority.

Sources: metrics.csv, model_summary.csv; summary figure 01 [2,5].

Higher recall has a false-positive tradeoff

FIXED THRESHOLDS PRESERVE THE DECLARED COMPARISON

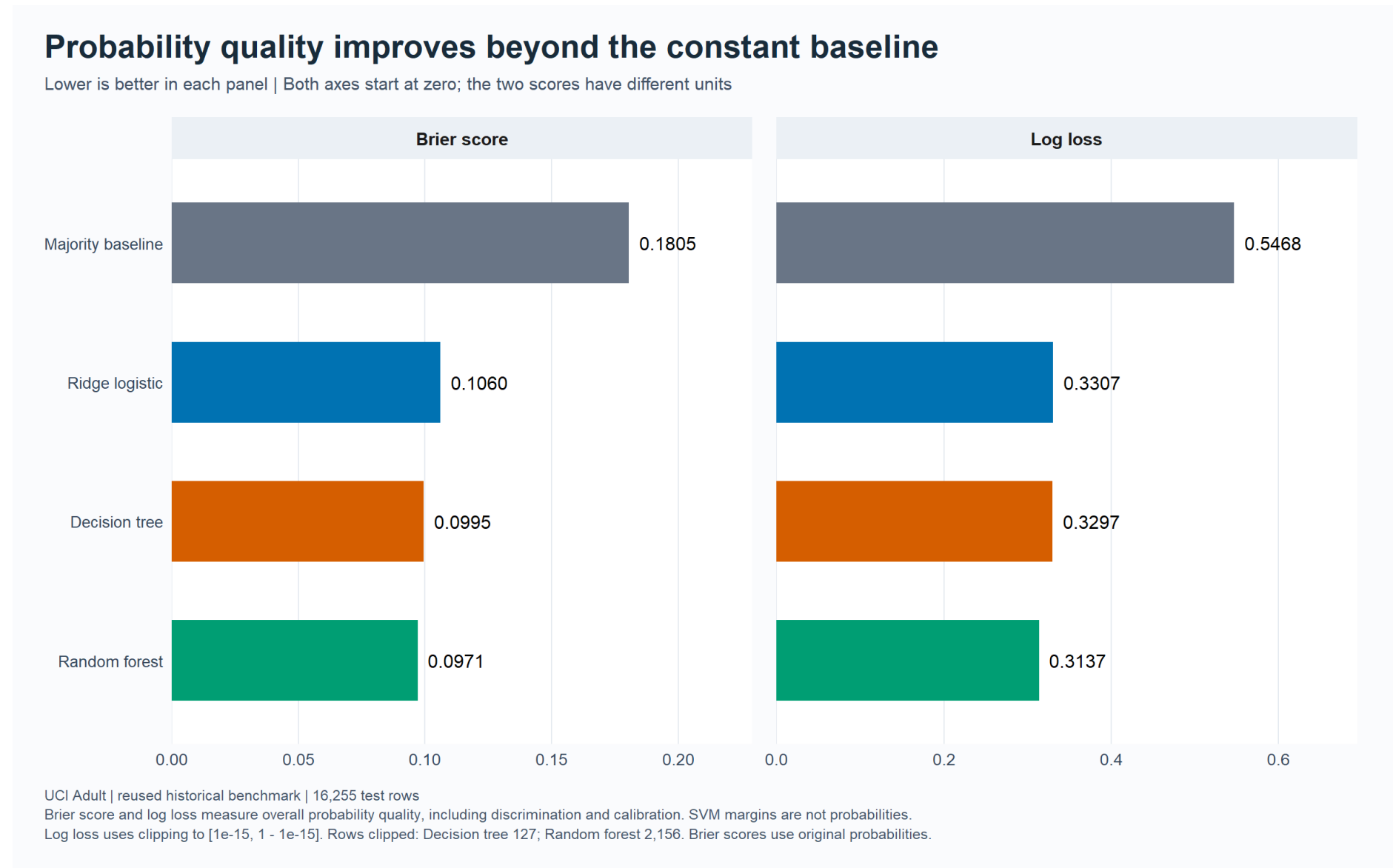


The tree makes 2,208 errors, 42.5% fewer than the baseline. Relative to the tree, the forest finds 23 additional positive records and adds 85 false positives, for 62 more errors overall.

Sources: metrics.csv, headline_statistics.csv; summary figure 02 [2,5].

Probability error improves beyond the baseline

Brier score and log loss answer a separate question



The forest has the lowest Brier score (0.0971) and log loss (0.3137). Log loss clips 2,156 forest probabilities and 127 tree probabilities at the declared numerical bounds.

Sources: metrics.csv, calibration.csv; summary figure 05 [2,5].

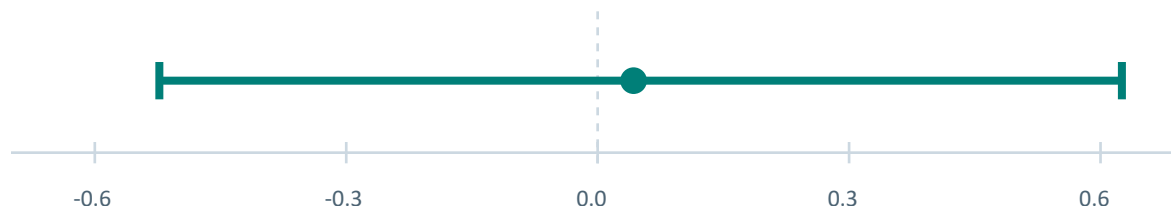
A narrow observed difference needs context

CONDITIONAL UNCERTAINTY AND METRIC DEFINITIONS

Tree minus forest: balanced accuracy

Observed difference: 0.043 percentage points

Conditional 95% paired interval: -0.523 to 0.626 percentage points.



Percentage-point difference; positive values favor the tree.

The interval includes zero. It does not establish a clear balanced-accuracy advantage, and it is not evidence of equivalence. The two estimates refer to the same evaluation records.

Bootstrap design. Resample 16,249 clusters defined by all 14 raw predictor fields, with replacement, for 1,000 paired replicates. Preserve all rows in each drawn cluster. Use the same draws for every model. Report percentile intervals for five metrics and all six trained-model pairs.

Scope. These intervals condition on the fitted models and reused test sample. They omit training and selection variability, previous test exposure, population shift, and unknown person-level dependence. Pairwise intervals are not adjusted for multiple comparisons.

Sources: paired_differences.csv, bootstrap_design.csv, calibration_summary.csv [2,5].

Read the metrics

Balanced accuracy

Mean of recall and specificity; gives both classes equal weight.

ROC AUC

Score ordering across positive and negative records, accounting for ties.

Average precision

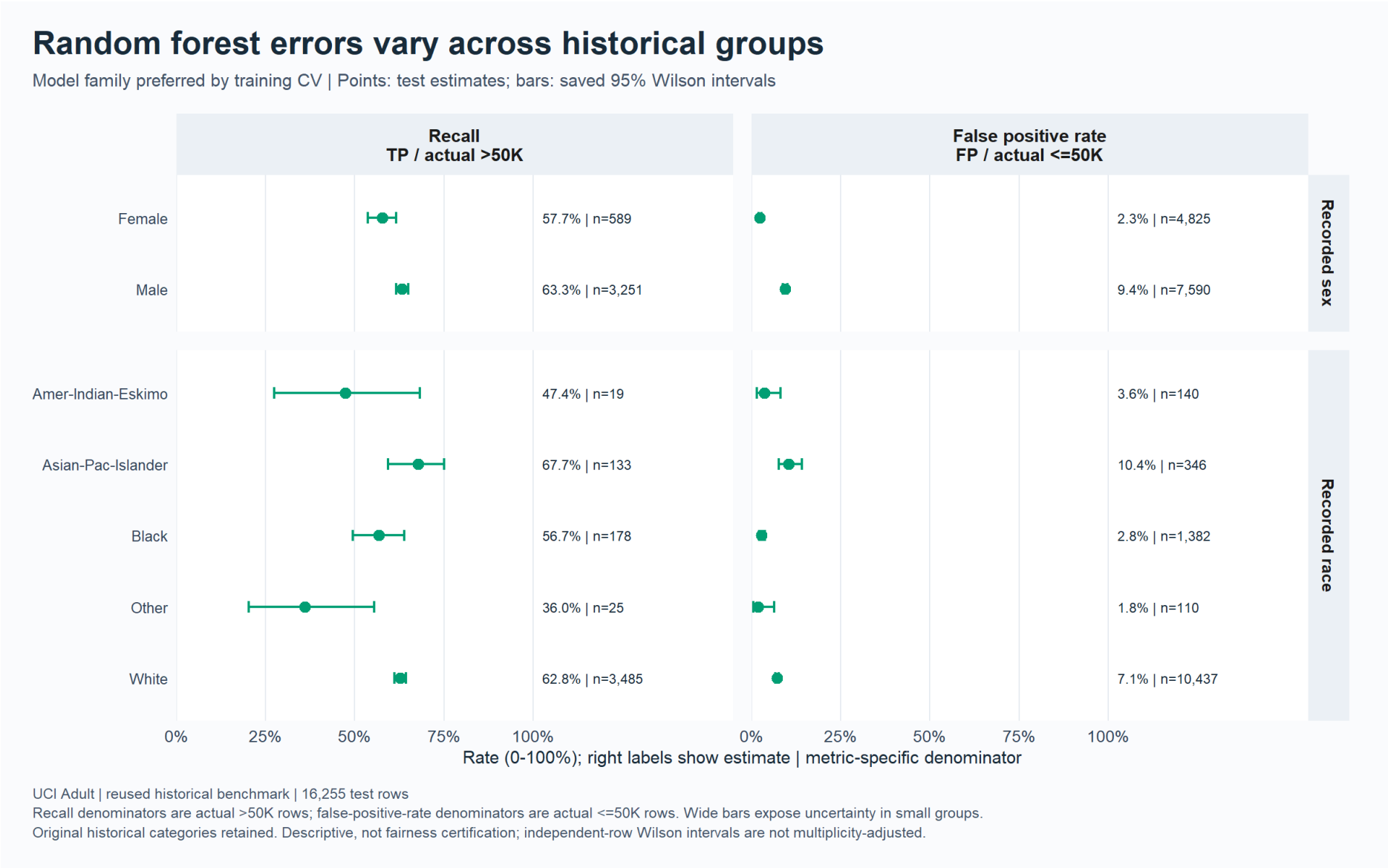
Stepwise integration over tied-score recall increments. Not trapezoidal PR area.

Calibration

Ten fixed probability bins describe reliability. Binned error depends on binning and does not rank overall model quality.

Group error rates require denominators

DESCRIPTIVE CHECKS FOR THE CV-SELECTED FOREST



The smallest positive denominators are 25 for Other and 19 for Amer-Indian-Eskimo, with 9 detected positives in each group. Wide intervals restrict comparisons.

Source: subgroup_metrics.csv and saved Wilson intervals; summary figure 08 [2,5].

Profile subsets have different case mixes

MODEL-INPUT PROFILES DO NOT IDENTIFY REPEATED PEOPLE

Test results differ between seen and novel model profiles

All five models | Same fitted models and fixed thresholds; two descriptive test subsets



UCI Adult | reused historical benchmark | 16,255 test rows

Seen means an identical model-input predictor profile occurs in training. It does not establish that rows identify the same person.

The subsets have different prevalence and case mixes. These point estimates do not isolate an overlap effect; no new intervals are inferred.

The forest has 72.12% balanced accuracy for 2,953 records with seen profiles and 79.23% for 13,302 records with novel profiles. Different case mixes prevent causal interpretation.

Sources: profile_overlap.csv, profile_sensitivity.csv; summary figure 09 [2,5].

Reproduce the workflow and verify the evidence

R 4.6.1 WITH 34 LOCKED PACKAGE VERSIONS

1. Restore and check

Use PowerShell from the repository root. The installer restores the recorded dependencies into the ignored local library. Installation and explicit downloads require internet access.

```
Rscript --vanilla scripts/install_dependencies.R
Rscript --vanilla tests/run_tests.R
Rscript --vanilla scripts/verify_evidence.R
```

2. Run the official benchmark

```
Rscript --vanilla IncomeLevelPrediction.R `
  --download --output results/v2-full
```

The backtick continues the PowerShell command on the next line. Use a new output directory: nonempty directories are refused. Defaults: seed 12345, five folds, 500 trees, and 1,000 bootstrap replicates.

For a smaller execution check, add **--smoke** and choose a separate output path. Smoke scores are not full-study results. Local inputs can be supplied with **--data-dir**; custom hashes prevent official-reference export.

3. Recreate the supplementary summaries

```
Rscript --vanilla scripts/create_results_summary.R `
  docs/evidence/v2-full-2026-10-08 `
  results/new-results-summary
```

Backticks continue this PowerShell command. It reads verified published aggregates and creates new charts and statistics. It does not train models, retune thresholds, or recompute intervals.

4. Check summary outputs

```
Rscript --vanilla tests/test_results_summary.R
Rscript --vanilla tests/test_results_summary_verifier.R
Rscript --vanilla scripts/verify_results_summary.R
```

Exact historical experiment. Its nine source/protocol/lockfile hashes are preserved under archive/v2-2026-10-08. For the complete original checkout, use publication commit **2ab72d12cf5290df1841da70b524740a70718c17** in a separate checkout. Current maintenance code has different hashes.

Publication boundary. Raw files, row-level predictions, fold IDs, fitted models, local libraries, and session details stay local. Public evidence uses an aggregate allowlist. Manifests verify bytes and provenance; they do not establish scientific validity by themselves.

Sources: README.md, renv.lock, verification scripts, VALIDATION.md [1,4,6].

Conclusions and interpretation limits

WHAT THIS PROJECT SUPPORTS

Supported conclusion	Practical interpretation
Training selection preferred the forest.	Mean training-fold balanced accuracy was 78.32%, versus 78.23% for the tree. This is the declared selection outcome, not independent performance confirmation.
Observed test criteria disagree.	The tree has the highest accuracy and balanced accuracy. The forest has the strongest observed ranking and lowest aggregate probability error. Their balanced-accuracy difference remains uncertain.
Useful ranking leaves classification errors.	Forest recall is 62.42%; it misses 1,443 of 3,840 positive records and generates 827 false positives at the fixed threshold. Error types matter alongside aggregate scores.
Diagnostics expose limits.	Subgroup denominators, calibration bins, profile subsets, conditional intervals, and preserved provenance show where apparently strong performance needs qualification.

Boundaries

Historical labels and sampling; unweighted metrics; reused test exposure; no person identifiers; conditional uncertainty; no fairness target; no causal or deployment claim.

Defensible portfolio value

An independently developed R study that evolved from a learning script into a transparent, reproducible evaluation with documented corrections and ggplot2 result communication.

A fresh external or temporal sample would be needed for stronger generalization evidence. Changes across versions are not controlled estimates of the effect of any one methodological revision.
Sources: report.md, methodology, validation record and derived statistics [1-5].

Evidence map and reference results

PUBLIC SOURCES AND ATTRIBUTION

Model	Accuracy	Bal. accuracy	Recall	ROC AUC	Avg. precision	Brier / log loss
Majority	76.38%	50.00%	0.00%	0.5000	0.2362	0.1805 / 0.5468
Ridge logistic	84.66%	75.79%	58.96%	0.8980	0.7354	0.1060 / 0.3307
Decision tree	86.42%	77.92%	61.82%	0.8910	0.7615	0.0995 / 0.3297
Random forest	86.04%	77.88%	62.42%	0.9090	0.7933	0.0971 / 0.3137
Linear SVM	84.80%	75.94%	59.14%	0.8968	0.7357	Not applicable

Full precision values remain in CSVs linked from sources [2] and [5]. SVM margins are uncalibrated, so probability metrics are not applicable. Baseline precision and MCC are undefined because it predicts no positive records.

Repository sources

- [1] [Project overview and usage](#)
- [2] [Published version 2 evidence](#)
- [3] [Declared version 2 protocol](#)
- [4] [Validation record](#)
- [5] [Additional ggplot2 results and statistics](#)
- [6] [Locked runtime and package versions](#)

Links are pinned to the merged source revision:
44a5f704aefa009647ccba0438b684cce52ef965

Data attribution

Becker, B. and Kohavi, R. (1996). *Adult*. UCI Machine Learning Repository.
[DOI: 10.24432/C5XW20](#)
[UCI Adult dataset and documentation](#)

Dataset license: [Creative Commons Attribution 4.0](#). The dataset license does not license this project's code; no code license has been selected.

Pages 3, 5-8, 10-11 reuse the project's original ggplot2 PNG exports with their captions and aspect ratios preserved. The interval diagram on page 9 reproduces saved endpoints without recomputing uncertainty.