

Income Level Prediction Using R

A reproducible comparison on the historical UCI Adult dataset

Esad Kopru

Revised analysis, October 2026

Study question and data

Predict the historical income class >50K

32,537

retained training records

16,255

retained benchmark test records

23.62% of test records belong to the positive class.

Historical, unweighted evaluation. Version 1 already used this test file.

Evaluation design

1 Preserve the official partitions

Audit duplicate records and cross-partition raw signatures.

2 Group five training folds

Fit imputation, encoding and scaling within each analysis fold.

3 Select by balanced accuracy

Compare 13 trained configurations plus the majority baseline.

4 Refit and evaluate fixed decisions

Use probability >0.5 or an oriented SVM margin >0 .

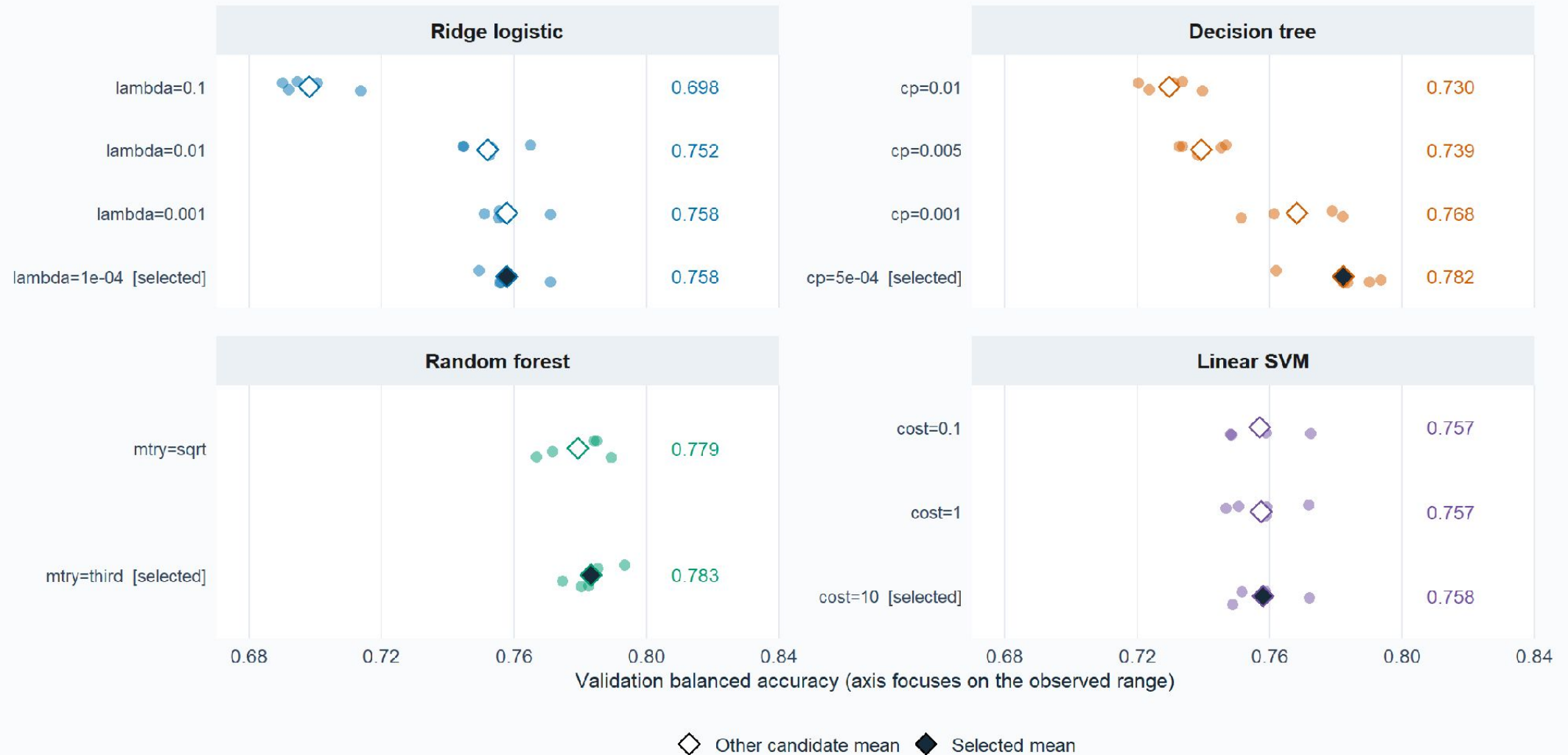
Models and selected settings

Model	Implementation	Selected setting
Majority baseline	Training prevalence	Constant prediction
Ridge logistic	glmnet	lambda = 0.0001
Decision tree	rpart	cp = 0.0005
Probability forest	ranger, 500 trees	mtry = floor(p / 3)
Linear SVM	e1071	cost = 10

Training CV preferred forest: 78.32% balanced accuracy vs 78.23% for tree.

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.

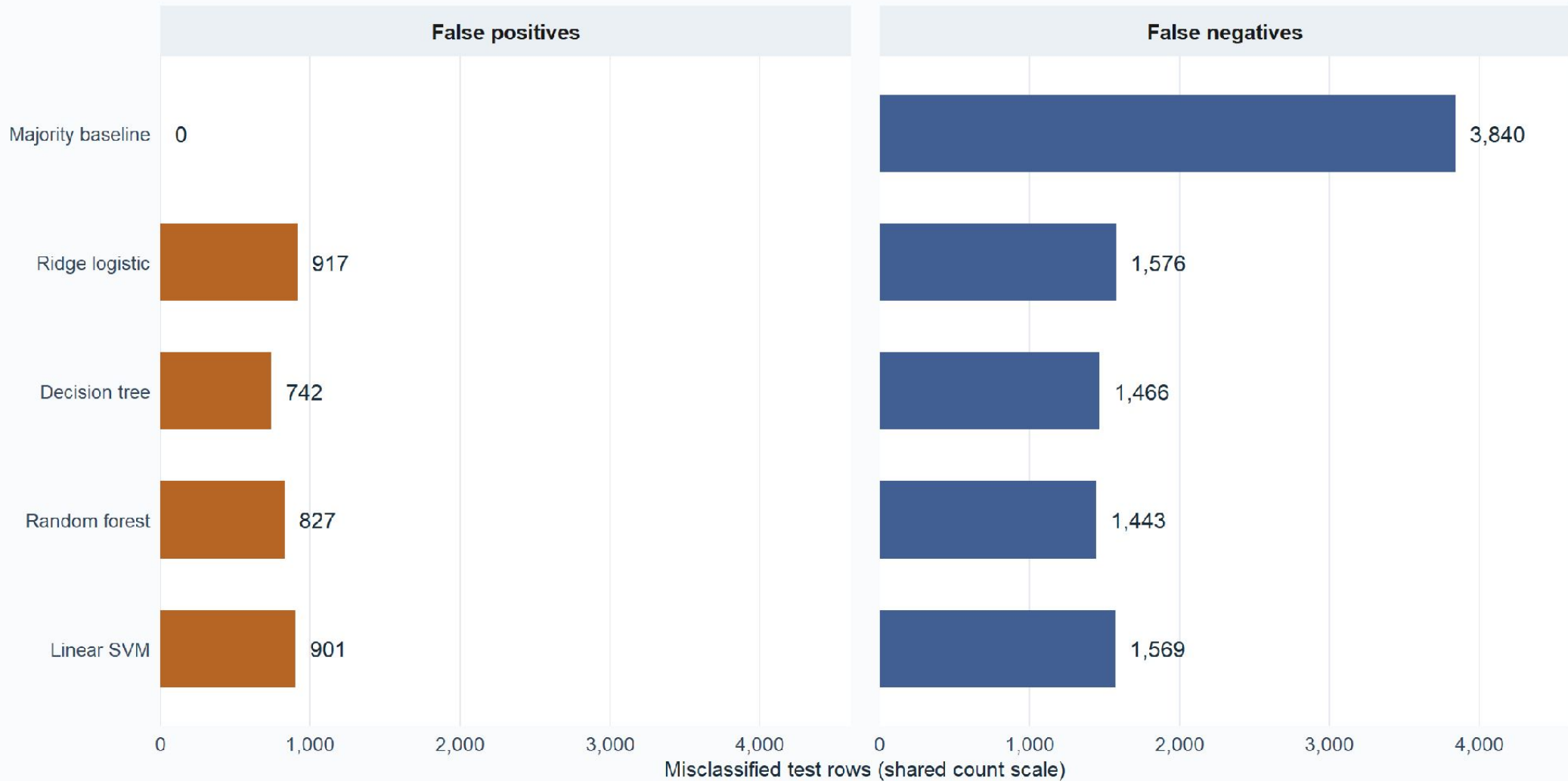
Performance depends on the metric

Model	Accuracy	Balanced accuracy	ROC AUC	Average precision
Majority	76.38%	50.00%	0.5000	0.2362
Ridge logistic	84.66%	75.79%	0.8980	0.7354
Decision tree	86.42%	77.92%	0.8910	0.7615
Random forest	86.04%	77.88%	0.9090	0.7933
Linear SVM	84.80%	75.94%	0.8968	0.7357

Balanced accuracy = (recall + specificity) / 2. AUC and AP measure ranking.

False negatives outnumber false positives in every model

False positive: $\leq 50K$ predicted $> 50K$ | False negative: $> 50K$ predicted $\leq 50K$



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

All five models use the same 16,255 test rows: 3,840 actual positives and 12,415 actual negatives. Counts are not error rates.

Decisions use the saved fixed thresholds. The majority baseline has zero false positives and misses every actual positive.

The forest changes the error balance

Compared with the decision tree

23

more positive records found

85

additional false positives

The forest still misses 1,443 of 3,840 positive records.

Fixed thresholds. No deployment cost model.

Probability quality improves beyond the constant baseline

Lower is better in each panel | Both axes start at zero; the two scores have different units



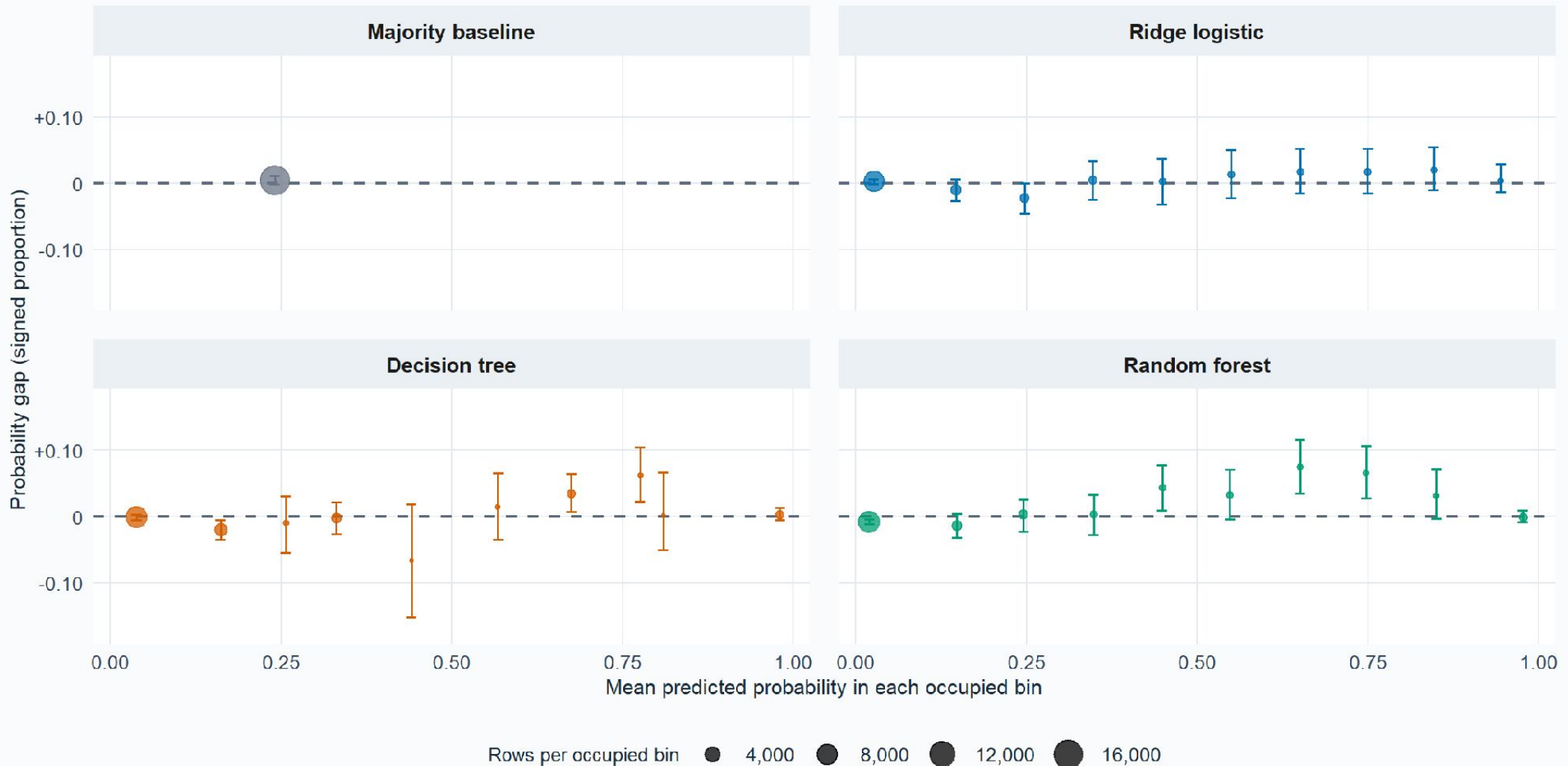
UCI Adult | reused historical benchmark | 16,255 test rows

Brier score and log loss measure overall probability quality, including discrimination and calibration. SVM margins are not probabilities.

Log loss uses clipping to $[1e-15, 1 - 1e-15]$. Rows clipped: Decision tree 127; Random forest 2,156. Brier scores use original probabilities.

Calibration gaps show overprediction and underprediction

Gap = mean predicted probability - observed >50K rate | Positive = overprediction; negative = underprediction



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

Ten fixed-width bins; empty bins omitted. Bars reflect saved 95% Wilson rate intervals; they assume independent rows and hold predictions fixed.

Zero marks agreement. Baseline agreement reflects prevalence, not discrimination. SVM excluded; no recalibration is fitted on this test set.

Conditional uncertainty limits the comparison

Tree minus forest, balanced accuracy

+0.043

percentage points

95% conditional interval: -0.523 to 0.626 points

1,000 paired cluster draws. Fixed models, reused test. Training and selection variability excluded. Zero lies inside the interval.

Random forest errors vary across historical groups

Model family preferred by training CV | Points: test estimates; bars: saved 95% Wilson intervals



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

Recall denominators are actual >50K rows; false-positive-rate denominators are actual <=50K rows. Wide bars expose uncertainty in small groups.

Original historical categories retained. Descriptive, not fairness certification; independent-row Wilson intervals are not multiplicity-adjusted.

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.

What the audit corrected

Record alignment

Consistent record handling replaces mismatched cleaned and raw row indices.

Evaluation leakage

Training folds learn preprocessing and select settings before test scoring.

Score interpretation

Correct score thresholds and continuous inputs support classification and ROC.

Reproducible comparison

Locked R packages, archived code and aggregate evidence preserve the analysis.

Conclusion

Training cross-validation selected the forest

The tree achieved 86.42% test accuracy. The forest led ranking with ROC AUC 0.9090 and average precision 0.7933.

The balanced-accuracy difference remains uncertain

Tree minus forest: +0.043 percentage points. The conditional 95% interval spans -0.523 to 0.626. It establishes neither a clear advantage nor equivalence.

Historical data. Reused test. Conditional uncertainty.

Sources and project resources

Public repository

github.com/efkopru/income-level-prediction

Methods, evidence and ggplot2 results

Versioned protocol, aggregate tables and downloadable figures

UCI Adult dataset

Becker and Kohavi (1996). DOI 10.24432/C5XW20

Data license: Creative Commons Attribution 4.0

Results shown throughout use the October 8, 2026 version 2 run.