

Improving Recall in Diabetes Risk Prediction

Class Imbalance and Feature Importance in BRFSS Health Data

Aarush Gowda

BRFSS 2015

84.6%

Positive-class recall

5,983 / 7,069

Positive cases detected

27.9%

Precision

0.822

ROC-AUC

BALANCED RANDOM FOREST | THRESHOLD 0.44 | HELD-OUT TEST SET: 50,736 OBSERVATIONS

RESEARCH QUESTION How do training-set balancing and classification thresholds affect recall, precision, and F2 in diabetes risk prediction?

01 Abstract & study data

High overall accuracy can hide missed positive cases when classes are imbalanced. This study compares Logistic Regression and Random Forest to examine how training-set balancing and classification thresholds affect detection of the combined prediabetes/diabetes class.

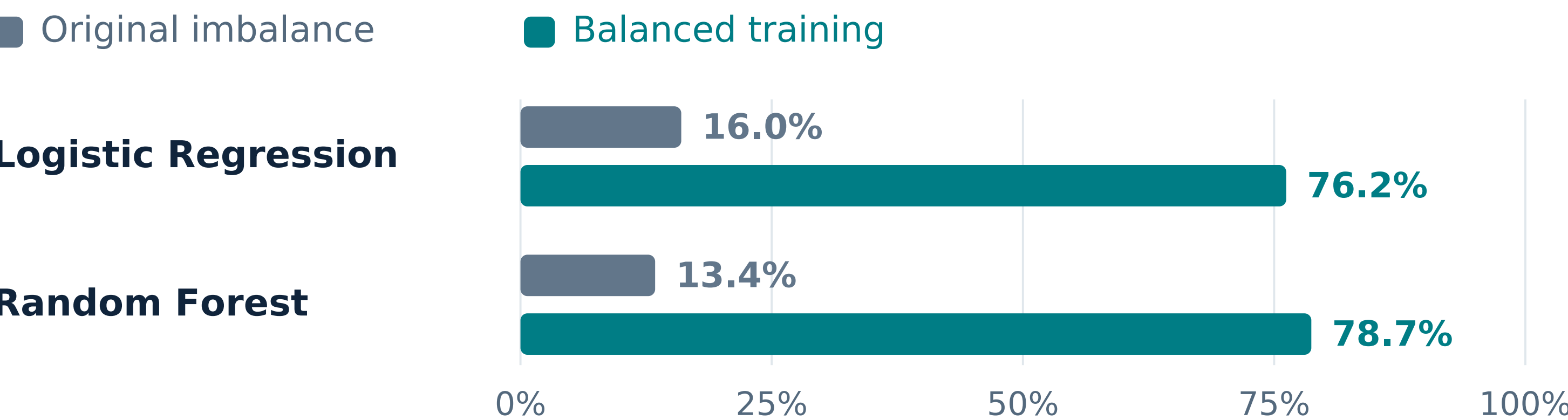
253,680 observations | 21 predictors

No diabetes: 86.07%

Prediabetes/diabetes: 13.93%

A Effect of training balance

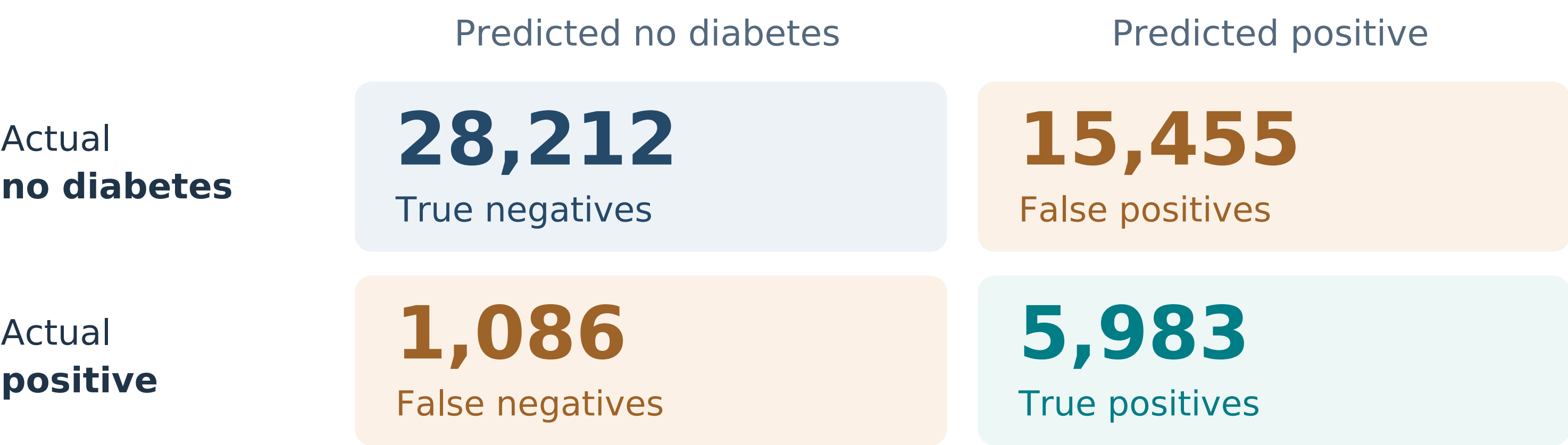
Recall at a common probability threshold of 0.50



A fixed cutoff isolates the effect of training-set balancing.

C Detection & false positives

Balanced Random Forest at threshold 0.44



Positive class = prediabetes/diabetes

02 Methods

- Split the data**
Stratified 60% train / 20% validation / 20% test; random seed 42.
- Balance training only**
Randomly undersample the majority class; preserve validation and test prevalence.
- Fit two classifiers**
300-tree Random Forest and scaled, L2-regularized Logistic Regression.
- Select probability cutoffs**
Maximize validation F2; keep the test set out of threshold selection.
- Evaluate and interpret**
Test-set metrics plus recall-based permutation importance on 20,000 test observations.

B Validation-selected thresholds

Test performance after maximizing validation F2

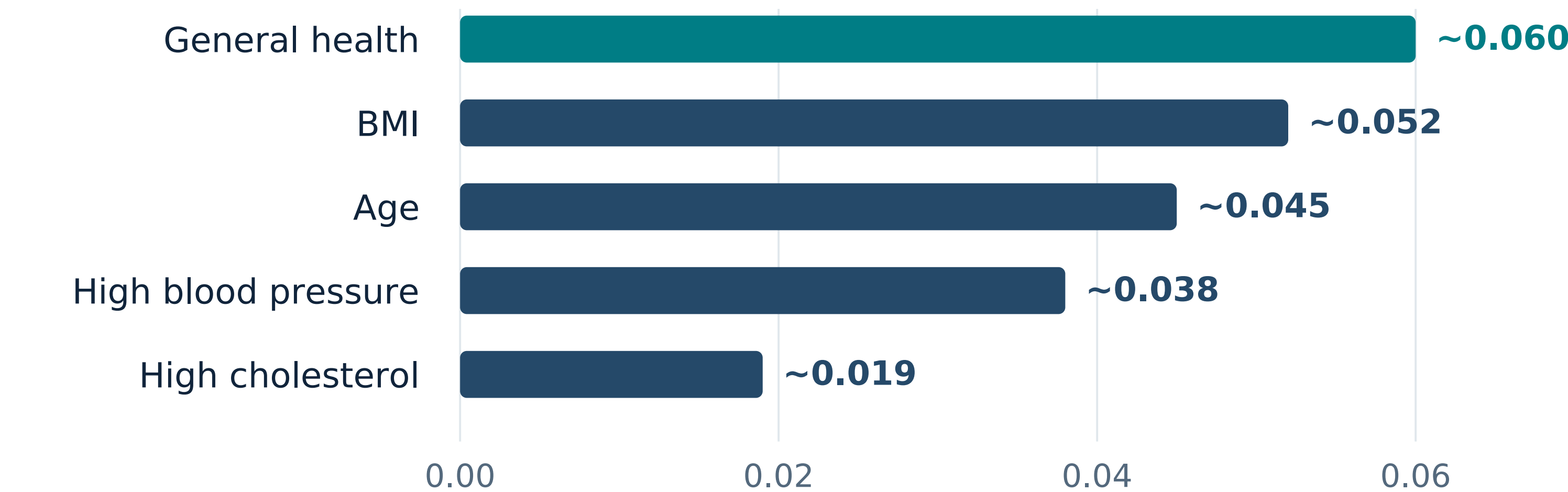
Model	Cutoff	Recall	Precision	F2
Balanced RF	0.44	0.846	0.279	0.602
Balanced LR	0.41	0.846	0.278	0.601
Imbalanced LR	0.11	0.829	0.283	0.598
Imbalanced RF	0.13	0.814	0.288	0.596

RF: Random Forest | LR: Logistic Regression

Similar F2 scores: balanced RF = 0.602; balanced LR = 0.601.

D Features supporting recall

Approximate mean recall decrease after feature shuffling



Five shuffles per feature on 20,000 test observations. Values are rounded from the paper; importance indicates association, not causation.

CONCLUSION

Balancing and threshold selection improved positive-case detection, with a substantial false-positive cost. Recall, precision, and F2 reveal tradeoffs that accuracy alone can hide. These models support screening research; they do not establish diagnoses.

LIMITATIONS & NEXT STEPS

Self-reported data; merged positive class; missing laboratory measures; no external or subgroup validation. Next: independent cohorts, subgroup assessment, and thresholds tied to screening costs.

SELECTED REFERENCES

CDC. BRFSS documentation (2025).
Teboul. Kaggle Diabetes Health Indicators (2021).
Pedregosa et al. JMLR 12 (2011), 2825-2830.

Recall: share of actual positive cases correctly detected.

Precision: share of positive predictions that are correct.

F2: combines precision and recall, giving recall more weight.