Ensemble machine learning for sex prediction of a worldwide craniometric dataset

Evan Muzzall[1], Chris Kennedy[1,2], and Aaron Culich[3]

[1] D-Lab [2] Division of Biostatistics [3] Berkeley Research Computing, Research IT

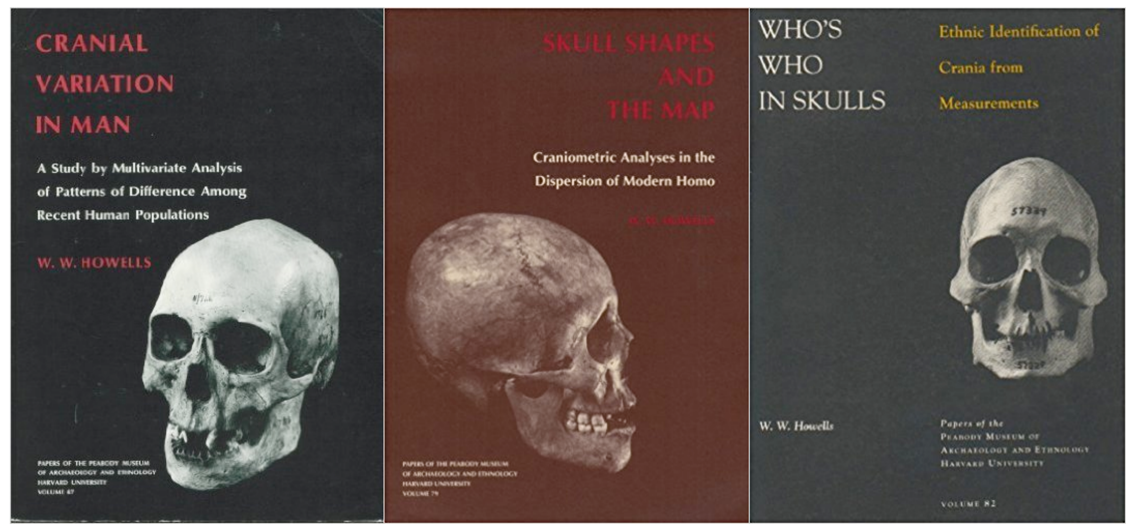
Introduction

-Sex estimation is an important first step for reconstructing the biological profile of unidentified human remains.

-Because girls and boys skeletally mature at different rates, correct sex estimation accuracy is important to prevent misidentification of other aspects such as age, geographic origin, etc.

-Human cranial metric diversity is a genetic proxy for human population structures & evolution and is used in the investigations of missing persons[1].

-**Objective:** As sex misclassification can lead to downstream errors in other variables, we demonstrate ensemble machine learning to capture relationships between macroscopic cranial sex estimation and craniometric variation.



Howells Craniometric Datasets

-Worldwide craniometric data collected by William W. Howells of the Harvard Peabody Museum[2,3].

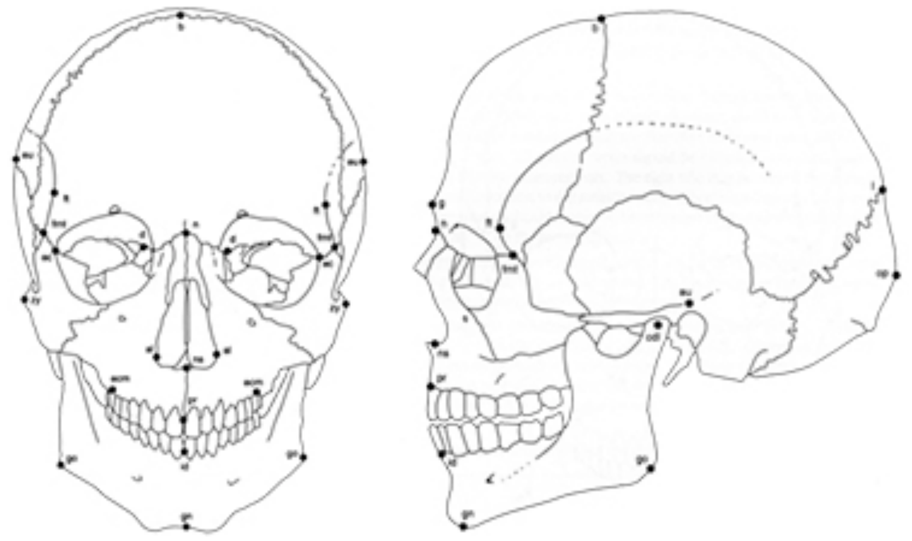
-**Training dataset:** 2,524 obs.

-**Test dataset:** 524 obs.

-**Covariates (82):** distances (mm) between cranial suture intersections and floating point landmarks.

-**Outcome:** binary female/male sex label determined by Howells largely via skull morphology (and others such as pelvic features when available).

-**Sex estimation:** Howells estimated sex using skull morphological features instead of more reliable methods such as from the pelvis.



https://osteoware.si.edu/guide/craniometrics

Methodology

-Apply supervised machine learning to human cranial metrics for binary female/male sex prediction.

-**Algorithms:** decision trees, random forest, bayesian additive trees, generalized additive models, multivariate adaptive regression splines, k-nearest neighbors, support vector machines, and gradient boosted machines

-**Benchmark:** standard main effects logistic regression

-**Ensembling:** SuperLearner-based weighted average of learners using cross-validated test set predictions.

-**Hyperparameter optimization:** explore sensitivity of each algorithm to hyperparameter tuning.

-**Performance:** cross-validated AUC and R-squared, along with nested cross-validation to evaluate ensemble performance.

-**Computation** was performed with the **Savio** cluster run by Berkeley Research Computing/Research IT.

Data preprocessing

-**Missing value imputation** using k-nearest neighbors.

-Standardized to mean 0 and standard deviation 1.

-Two **missingness indicators** added.

-Confirmed design matrix was full-rank using QR decomposition.

-These same steps for training preprocessing were then applied to the **test dataset** using the training data parameters.

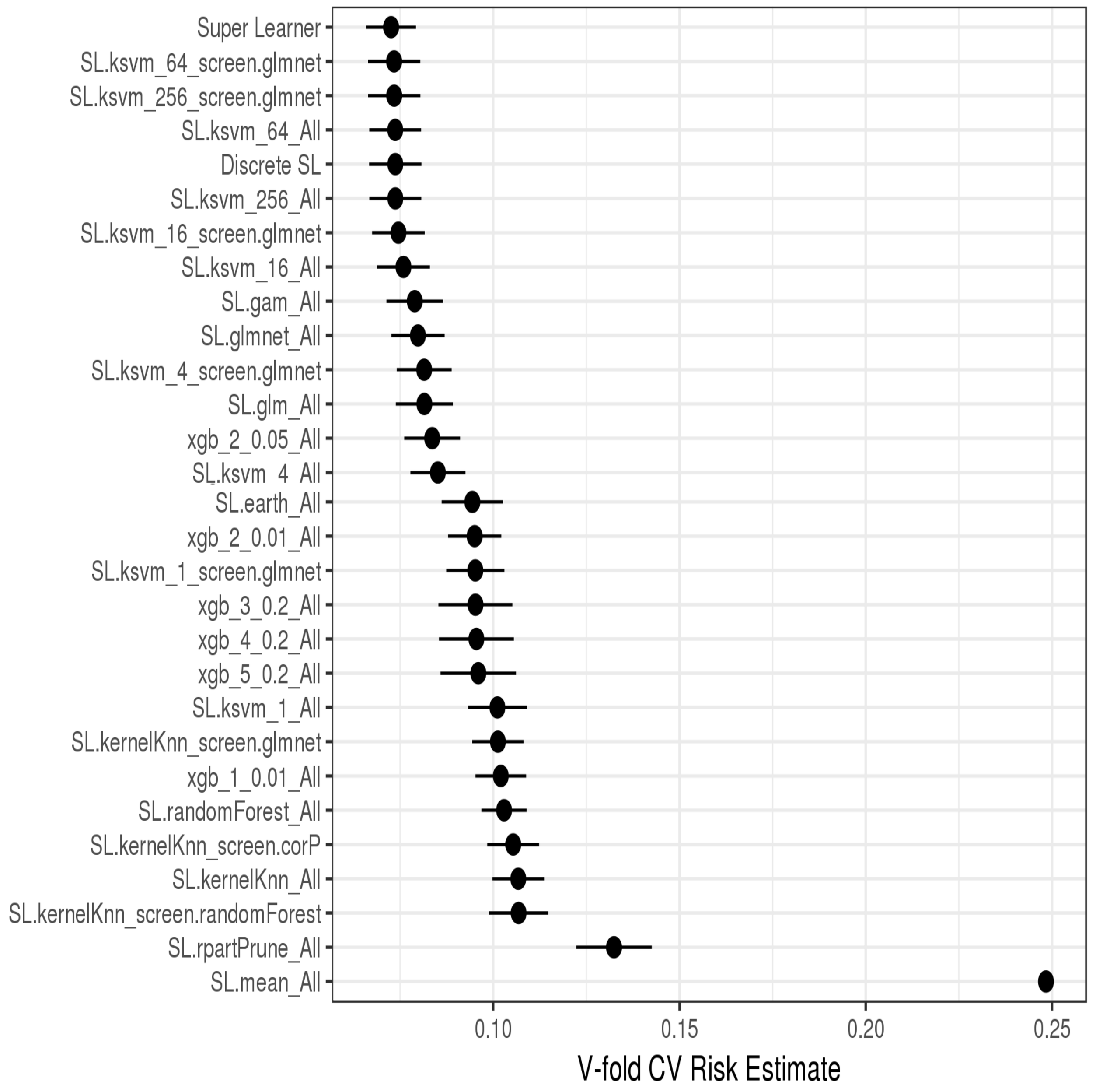
Hyperparameter optimization example: Support Vector Machines

	Library	Learners	AUC	95% C.I.
1	kSVM tune C, kernel, screener	37	0.966	0.960 – 0.972
2	kSVM tune C, kernel	15	0.965	0.959 – 0.971
3	kSVM tune C	8	0.962	0.955 – 0.968
4	kSVM default	2	0.959	0.952 – 0.966
5	SVM default	2	0.937	0.927 – 0.946

-The best performance is achieved with a SuperLearner ensemble of 37 SVMs, but next best is almost as good.

-Configurations vary by C (regularization), kernel, and feature screening (lasso).

Ensembling results (abridged*)



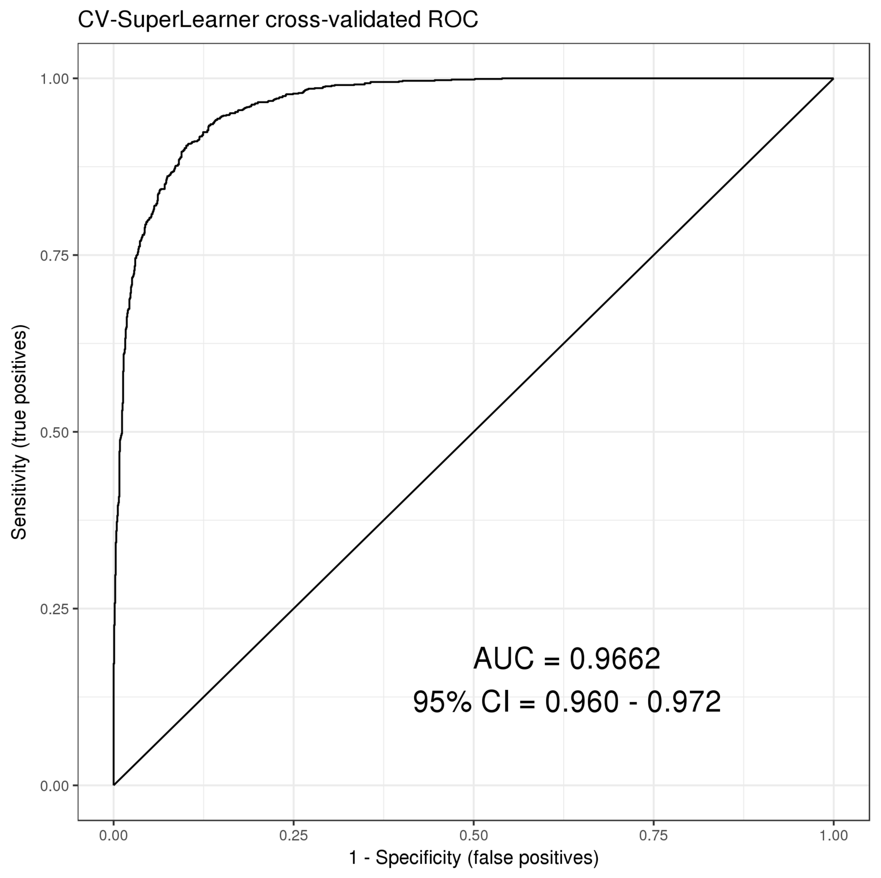
*Some configurations of XGBoost removed from chart to simplify presentation.

Cross-validated training and test set performance

Method	CV Training AUC	Test AUC
Logistic benchmark	0.956	0.797
Best single learner (SVM)	0.965	0.912
SuperLearner	0.966	0.910

(Paper will include CIs for easier comparison.)

Ensemble CV ROC on Training Set



Discussion

- Machine learning ensembles provide a huge improvement over logistic regression for sex-prediction using Howells' craniometric data.
- Extensive hyperparameter optimization for craniometric sex-prediction, beyond prior studies.
- This approach yields more accurate results that can improve identification of missing persons and understandings of human population variation.

Limitations

- **Measurement error** in outcome and covariates documented by Howells himself.
- Howells reviewed individuals with **high residuals** from the regression analysis and sometimes reclassified them, likely biasing results in favor of linear regression especially on training set.
- **Test dataset** contains wider variety of sources than the training dataset[2].

Future directions

- How to ensure minimal-error sex identification (labeling) in order to optimize machine learning's potential?
- Can museums that house skulls in the Howells dataset digitize them so covariates can be recalculated from these new digital models?
- Use CT scans to create new datasets comprised of living people of known age and sex?
- Employ deep learning on these image-based projects to more comprehensively examine sex prediction, population structure, and missing persons.

References

1. Buikstra JE, Ubelaker DH. 1994. Standards for data collection from human skeletal remains. Field Museum of Natural History.
2. Howells WW. Notes and comments: Howells' craniometric data on the internet. Am J Phys Anthropol 101:441-442.
3. Polley E, LeDell E, Kennedy C, Lendle S, van der Laan M. 2016. SuperLearner R package. CRAN.

Acknowledgements: We thank Janet Torres. Dr. Benjamin M. Auerbach maintains these data here: https://web.utk.edu/~auerbach/HOWL.htm