Predicting Age based on Bio-measurements

Peyton Ardoin, Khalil El-abbassi, Aditya Baisakh, Marilin Guerrero Laos, Tre' Henson, Matthew Lemoine, Austin Louque, Efren Mesino Espinosa, Aduragbemi Oyerinde, Shalini Shalini, and Gowri Priya Sunkara

Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana
Faculty Advisors: Dr. Nadejda Drenska and Dr. Peter R. Wolenski

Introduction

This is a continuation of the Age prediction project. Our motivation is to predict a given patient's biological age. This can provide great insight into their relative health for their age. In this poster, we explore the results of Multiple Linear Regression, XGBoost, and Support Vector Regression using the Caesar Dataset.

Dataset Cleaning

- The original dataset contained over 120 variables per patient. After removing non-essential and redundant features, 94 anthropometric biomarkers were retained.
- Patients with incomplete records were excluded, removing 227 subjects and leaving 2,164 complete records for model training and evaluation.
- The *Bust/Chest Circumference Under Bust (mm)* measurement was systematically missing for male patients, so six dataset configurations (combined, male-only, and female-only, each with and without the under-bust feature) were created to avoid missing-value issues and support gender-specific analysis.

Multiple Linear Regression with PCA Results

We predict chronological age from anthropometric biomarkers using segmented Multiple Linear Regression (MLR) via Correlation Analysis and Principal Component Analysis (PCA)

- Preprocess anthropometric biomarkers using correlation filtering, VIF-based multicollinearity reduction, feature standardization, and PCA (95% variance retained) within each age group.
- Train age-specific Multiple Linear Regression (MLR) models and evaluate performance using MAE, RMSE, and R^2.
- Correlation analysis, VIF, and PCA reduce feature redundancy and multicollinearity, improving model robustness and predictive stability.

Age Group	Samples	No. of PCs	Variance	RMSE
0–20	58	17	95.36%	0.56
20–30	539	27	95.20%	2.69
30–40	612	26	95.01%	2.88
40–50	639	27	95.11%	2.87
50–60	396	28	95.05%	2.86
60–70	142	25	95.29%	1.85

- The final segmented MLR model achieved an MAE of 2.331 years, an RMSE of 2.746 years, an MSE of 7.538 years^2, and an R^2 of 0.9486, demonstrating strong predictive accuracy for biological age estimation.
- PCA reduced the dimensionality of the biomarker space while retaining approximately 95% of the variance within each age group.
- This pipeline achieved an overall RMSE of 2.75 years and an R^2 of 0.9486, substantially improving upon the initial performance.

MLR with PCA Results

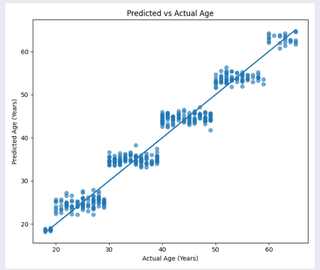

Figure 1. Segmented MLR w/ PCA Model; RMSE = 2.75 years

- This figure compares the predicted biological age with the actual chronological age, where points closer to the diagonal line indicate more accurate predictions.

XGBoost

Improvements over the original 12.88 RMSE global prediction error from first dataset due to:

- Caesar dataset having 4 times as many patients
- Splitting sexes removes some physical discrepancy noise that was unrelated to age-based changes
- Hyperparameters tuned via 5-fold CV

Model	Error
XGBoost w/ tuned hyperparameters (Males)	**8.24 years**
XGBoost w/ tuned hyperparameters (Females)	**8.66 years**

Table 1. Results from XGBoost tests.

Support Vector Regression

- Support Vector Regression (SVR) predicts continuous values by fitting an optimal function within an ϵ-insensitive margin, where only support vectors influence the model.
- Hyperparameters (C and ϵ) are optimized using random search, and SHAP values are used to identify and retain the most important biomarkers.
- The final SVR model is combined with K-Means clustering, prediction correction to reduce regression dilution, and centile deviation analysis to improve predictions and identify atypical subjects.

Input	Result
Whole data set split into 10-year age columns	7.9 years
SHAP tuned model (0.02)	2.4 years
SHAP tuned model with tuned parameters	1.94 years
SHAP tuned model with tuned parameters and tuned SHAP thresholds	1.92 years
New data set with SHAP and K-Means on one age bracket	8.54 years

Table 2. Results of SVR

SVR SHAP with K-Means with RMSE=8.54 years

Highlighted below are some of the plots from the recent SVR runs

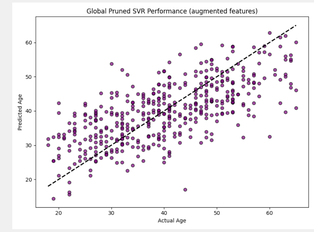

Figure 2. SVR SHAP with K-Means; RMSE=8.54 years

Age Gap versus Centile Deviation

Figure 3. Age gap Vs. centile deviation (Left), Probability density of Age Gap (Right)

Table 3. Results from SVR run output to CSV

PPT ID	Actual Age	Corrected Gap	Centile Deviation
478	42 years	-1.5 years	-13.9
1745	51 years	-11.3 years	-24.8
1851	64 years	-9.1 years	-11.9

Acknowledgements

- We would like to thank Dr. Steven B. Heymsfield from Pennington Biomedical Research Center for giving us the datasets and supporting the evaluation of the models.
- We would like to thank Prof. Peter R Wolenski and Dr. Nadejda Drenska for their help with this project and their guidance.

References

[1] G. Agrahari, K. Bist, M. Pandey, J. Kapita, Z. James, J. Knox, S. Heymsfield, S. Ramirez, P. Wolenski, and N. Drenska, "Predicting anthropometric body composition variables using 3d optical imaging and machine learning," *Frontiers in Bioinformatics*, vol. 6, p. 1722578, 2026.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[3] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, pp. 293–300, 1999.

[4] C. A. Fermín-Martínez, A. Márquez-Salinas, E. C. Guerra, L. Zavala-Romero, N. E. Antonio-Villa, L. Fernández-Chirino, E. Sandoval-Colin, D. A. Barquera-Guevara, A. Campos Muñoz, A. Vargas-Vázquez, C. D. Paz-Cabrera, D. Ramírez-García, L. M. Gutiérrez-Robledo, and O. Y. Bello-Chavolla, "Anthropoage, a novel approach to integrate body composition into the estimation of biological age," *Aging Cell*, vol. 22, no. 1, p. e13756, 2023.