EDLD 654 Final Project: BrainAGE in Adolescence

Lucy Whitmore[1]

[1] University of Oregon

EDLD 654 Final Project: BrainAGE in Adolescence

All materials and scripts are available at https://github.com/LucyWhitmore/EDLD-654

Research Problem

As both neuroimaging and computational methods have improved in recent years, there has been interest in using machine learning methods to identify deviations in normative brain aging and development. One proposed method is the Brain-Age Gap Estimation (BrainAGE), which quantifies the difference between one's chronological age and their age as predicted by machine learning models trained on neuroimaging data, often structural MRI measures (Brown et al., 2012). In older adult populations a positive BrainAGE, or a predicted age older than someone's chronological age, has been interpreted as reflecting premature brain aging, and has been associated with risk for Alzheimer's disease and cognitive decline (Franke & Gaser, 2019). More recently, there has been interest in applying BrainAGE models to adolescent populations, where it has been hypothesized that BrainAGE could be related to risk for psychopathologies such as anxiety and depression. However, the majority of BrainAGE models are either trained exclusively on data from adult populations, or use dataframe from across the lifespan, but are not clear about how many adolescents are actually included in the model training.

Creating a BrainAGE model specifically for adolescent populations could improve the quality of predictions, better enabling researchers to accurately predict age. By improving BrainAGE predictions, we can also establish which outcomes and processes are related to BrainAGE in adolescence, both improving our understanding of development, and providing a possible indicator of risk.

Data Description

The data used in the creation of the following models were generated for use in this project, and are designed to match the format and distributions of data from the Adolescent Brain and Cognitive Development Study, a longitudinal multisite study of nearly 11,000 adolescents from the US, who are followed for 10 years. Along with other activities, participants take part in an MRI scan every two years. Currently, data are available from the first two waves of data collection. For this project, simulated data were created using the sim_df() function in R. 10,000 observations were simulated from data from the first two waves of the ABCD study. Simulated data were generated from a normal distribution, using the same distributions and correlations as the original data.

The data consists of 173 columns and 10,000 rows. One column represents age, expressed in months, which will be used as the outcome variable. The predictors are 172 numeric columns, each representing a volume or area measurement from a specific brain region. Column names containing "*vol*" represent volume measurements and columns containing "*area*" represent area measurements. 104 columns represent volume measurements, and 68 represent area measurements. For the age column (outcome variable), the range of values is 6.08-15.19 years and the mean is 10.9 years.

All predictors are continuous variables. As the data were simulated, there were no missing values. During data preparation, predictors with zero or near-zero variance were removed, and all predictors were standardized. The data were split into training and testing samples using an 80/20 split, resulting in 8000 observations in the training set and 2000 observations in the test set.

Model Description

To determine which model type provides the best fit, three different modeling approaches were used. These models included an unregularized linear regression, a linear

regression with a lasso penalty, and a bagged tree. All three models were constructed using 10-fold cross-validation, and were evaluated using MAE, RSQ, and RMSE, as these metrics are commonly used in the BrainAGE literature. Hyperparameter tuning procedures for each model are described below. I used R [Version 4.1.1; R Core Team (2021)][1],

for all the analyses.

Model 1 - Unregularized Linear Regression

No hyperparameters were tuned for the unregularized linear regression.

Model 2 - Linear Regression with LASSO Penalty

For model 2, the hyperparameter alpha was set to 1, and lambda was tuned with values from 0 to 0.015, in intervals of .001.

Model 3 - Bagged Tree

For model 3, hyperparameter mtry was set to 172 (the number of predictors), min.node.size was set to 2, and max.depth was set to 60. The num.trees hyperparameter was tuned using values 5, then a sequence from 20 to 200 in increments of 20.

Model Fit

The best fitting model was unregularized linear regression, followed by LASSO regression, then bagged trees. Individual model fits are described below, and shown in Table 1. As the outcome (age) was continuous, no cutoff point was needed. All reported performance metrics are based on performance on the test set.

---

[1] We, furthermore, used the R-packages *caret* [Version 6.0.90; Kuhn (2021)], *dplyr* [Version 1.1.2; Wickham, François, Henry, Müller, and Vaughan (2023)], *faux* [Version 1.2.1; DeBruine (2023)], *ggplot2* [Version 3.4.2; Wickham (2016)], *papaja* [Version 0.1.0.9997; Aust and Barth (2020)], *recipes* [Version 1.0.6; Kuhn, Wickham, and Hvitfeldt (2023); Kuhn et al. (2023)], and *vip* [Version 0.4.1; Greenwell and Boehmke (2020)].

Model 1 - Unregularized Linear Regression

Model 1 (unregularized linear regression) performed with an RSQ of 0.39, MAE of 0.79, and RMSE of 0.97 on the test set.

Model 2 - Linear Regression with LASSO Penalty

Model 2 (unregularized linear regression) performed with an RSQ of 0.36, MAE of 0.79, and RMSE of 0.97 on the test set. The best lambda value was 0.001.

Model 3 - Bagged Tree

Model 3 (bagged tree) performed with an RSQ of 0.22, MAE of 0.88, and RMSE of 1.08. The best value for num.trees was 200.

Model	RSQ	MAE	RMSE
Logistic Regression	0.39	0.79	0.97
Logistic Regression with LASSO Penalty	0.36	0.79	0.97
Bagged Trees	0.22	0.88	1.08

Figure 1. Model performance comparison

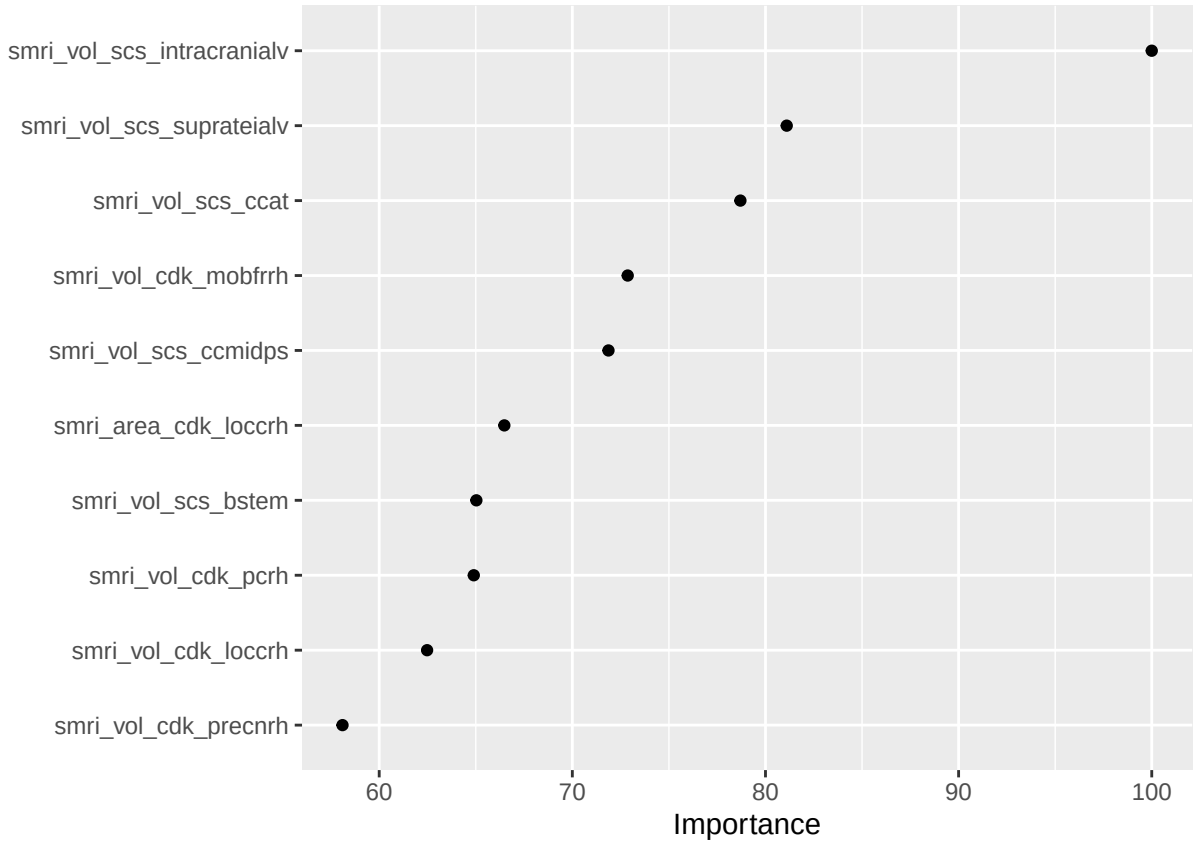

Figure 2. Variable importance of best-fitting model.

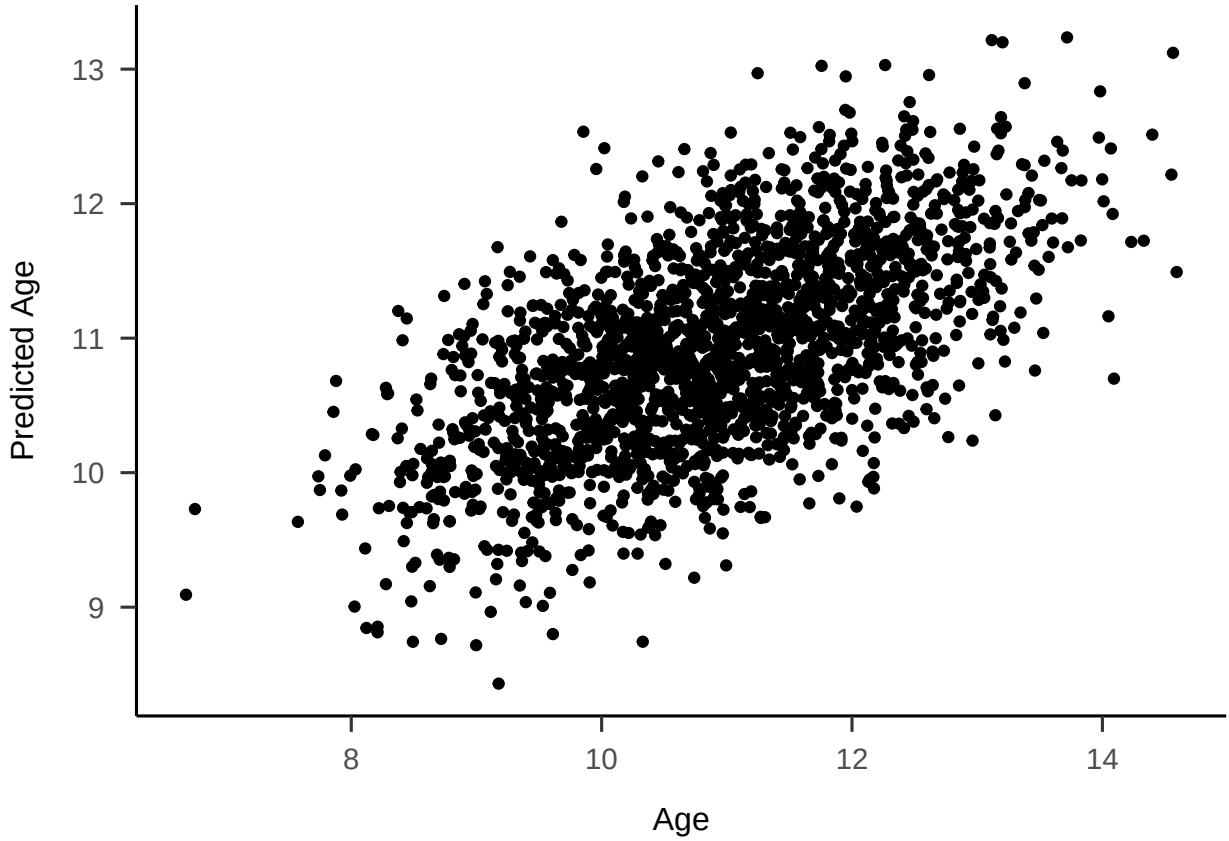

Figure 3. Predictions versus actual values for best-fitting model.

Discussion and Conclusion

Overall, the best-fitting model for predicting BrainAGE from structural MRI features was an unregularized linear regression. However, a LASSO regression approach resulted in a nearly identical performance, which was unexpected, but not impossible. The more surprising finding was that a bagged tree approach resulted in the poorest model fit out of the three models tested. Generally, these models perform similarly, or potentially better than a simple unregularized regression, and it's unclear why the bagged tree performed poorly. One potential explanation could be inadequate hyperparameter tuning for the bagged tree. Some hyperparameters were set explicilty rather than being tuned, including mtry, max.depth, and min.node.size. Tuning these hyperparameters, or simply picking different values could potentially have improved performance. Additionally, the best value of num.trees was the maximum value tested, and it's possible that an even larger value would have improved the

model. In the future, I would want to test more complicated models that allow for additional hyperparameter tuning, including random forest and gradient boosting models.

In terms of predictors, some of the highest contributing predictors were intracranial volume, supratentorial volume, anterior corpus callosum volume, right medial orbitofrontal volume, mid-posterior corpus callosum volume, right lateral occipital area, brainstem volume, right precuneus volume, right lateral occipital volume, and right precentral volume. The top predictor (intracranial volume) actually indicates one common issue with BrainAGE models. Intracranial volume refers to total brain size, and indicates that the model considers overall brain size as an important factor in determining age. However, one potential issue with that interpretation is that there are large individual differences in brain size, and overall brain size doesn't change as much during adolescence as certain areas redistribute, or gray matter is displaced by white matter due to myelination and gyrification, the process of the brain forming deeper/more extensive folds. Of the remaining top predictors, measurements related to the corpus callosum and brainstem replicate findings from previous BrainAGE models, which have determined those areas as high contributors.

In summary, adolescent BrainAGE was best predicted using an unregularized linear regression, though LASSO regression performed almost identically. Bagged trees performed worse than linear regression, but it's probable that more advanced tuning or models such as gradient boosted trees could improve performance. For my own work, this finding has led me to think more about the potentials pros and cons of complicated models. Many BrainAGE models use fairly advanced methods, such as extreme gradient boosting. While it's likely those models will still perform better than an unregularized regression, I'm more curious about how much of an improvement they actually provide. Additionally, I've been thinking much more about the trade-off between performance benefits and increased difficulty in communicating methodology, and I'm interested in seeing whether the performance benefits of more complicated models are worth the prospect of researchers running models that they

don't really understand, and may not be able to explain to an audience.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Brown, T. T., Kuperman, J. M., Chung, Y., Erhart, M., McCabe, C., Hagler, D. J., ... others. (2012). Neuroanatomical assessment of biological maturity. *Current Biology*, *22*(18), 1693–1698.

DeBruine, L. (2023). *Faux: Simulation for factorial designs*. Zenodo. https://doi.org/10.5281/zenodo.2669586

Franke, K., & Gaser, C. (2019). Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, *10*.

Greenwell, B. M., & Boehmke, B. C. (2020). Variable importance plots—an introduction to the vip package. *The R Journal*, *12*(1), 343–366. Retrieved from https://doi.org/10.32614/RJ-2020-013

Kuhn, M. (2021). *Caret: Classification and regression training*. Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., Wickham, H., & Hvitfeldt, E. (2023). *Recipes: Preprocessing and feature engineering steps for modeling*. Retrieved from https://CRAN.R-project.org/package=recipes

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr