James Hopham
BST 432
December 17, 2025

DNA Methylation Age Prediction

Abstract

DNA methylation is a genomic process used to regulate the expression of a given gene. The application of methylation clock models to predict age can be used to address health assessments as well as other areas of interest such as cancer research; however, these are challenging to build due to the vast amount of genetic information that can be attributed to a given individual. As a result, we must use methods that are able to address the curse of dimensionality that arises due to the nature of this high-dimensional problem. For this project, I employ elastic net, boosted trees and random forest as potential avenues to accurately model the relationship between a collection of approximately 500,000 CpG sites and the age of a given individual, and compare the strengths and weaknesses of each approach for their success as a methylation clock. Each of the proposed approaches demonstrated an RMSE of less than 4 on both the training and testing sets of the data.

Introduction

As a human develops from an embryo, to an infant, and then to adulthood, the expression of DNA changes throughout the cycle of one's life. To have this type of specialized development, mechanisms to regulate the expression of certain genes is necessary. DNA methylation is a regulatory process where a methyl group is bound to a nucleotide on a strand of DNA through the help of DNA methyltransferase. Once the methyl group is bound, it impacts the expression of a given gene by acting as a physical blockage, or by recruiting domain binding proteins that further recruit correpressors [1].

DNA is used inside cells as a script of information that provides the basic coded instructions on how to build important necessities for many functions. DNA is transcribed by RNA polymerase II into mRNA which is further translated into tRNA to build proteins. In order for DNA to undergo the process of transcription, certain regions play a vital role in allowing the process to occur including promoters, enhancers, and silencers. If, for example, the promoter region upstream of a given gene is physically blocked, RNA polymerase will not be able to bind to that region, and the corresponding gene will not be transcribed.

Another consequence is that transcription factors that provide key contributions to efficiency will not be able to provide support. Methylation also can directly lead to compaction of the DNA into heterochromatin, further closing the region off from being accessed.

Methylation switches genes on and off depending on whether they bind to a region that either promotes transcription or represses it. As a result, this is a key regulatory process that facilitates changes in expression in various areas including cell differentiation, brain development, and other key functions. It is expected for certain methylation to patterns occur throughout the life of an individual. As a result, the idea of methylation clocks can be applied where given the levels of methylation or demethylation found in a patient, an estimate of the person's age can be determined.

This further can be thought of in terms of epigenetics, where impacts of environmental, lifestyle, and other daily exposures or habits can contribute to direct effects on genes themselves. For example, a diet that consists largely of fried foods may introduce free radicals into the body that can directly affect mechanisms related to methylation which ultimately affects gene expression. The concept can then be further extended to assessing the difference between a person's age against their predicted biological age determined by a methylation clock to assess if there are any signs of premature aging. Other applications include the analysis of cancer cells, which demonstrate abnormal methylation patterns, and the analysis of differentiated somatic cells that are turned into induced pluripotent stem cells by Yamanaka factors, which demonstrate the same methylation age as embryonic stem cells.

This area of research, however, is complicated by the incredible number of genes that must be considered as a part of the model discovery process. Different sequencing arrays range in the number of methylated sites that are reported and range from 20,000 genes to 480,000 to 900,000 based on the given machine employed for the study. The large number of features ultimately gives rise to a high-dimensional problem where the number of features greatly outnumbers the number of samples.

When p>>n, we must consider the curse of dimensionality. This setting is challenging as models become less stable and are muddled by increased noise, difficulty in detecting signal, and computationally expensive operations. We must also additionally consider the attributes and behavior of DNA and the mechanisms that act upon genes themselves in terms of its ramifications on quantitative modeling which has ramifications on linear and nonlinear interactions and data processing.

Many approaches have been employed to tackle this high-dimensional task including elastic net as seen in Horvath’s clock, the addition of phenotypic clocks by Hannum, causal inference among genes in CausAge, and neural networks in AltumAge [3]. In 2013, Horvath employed an elastic net approach to model age against a processed set of around 21,000 genes which ultimately identified around 330 genes that potentially contribute significantly to the final prediction [5]. Hannum followed by integrating phenotype data into the model to add further context to learn from. AltumAge employed neural networks and additionally found that some of the genes originally identified by Horvath also demonstrated nonlinear interactions.

In this project, I attempt to propose 3 approaches for building methylation clocks on datasets collected from five different studies with the total combined data containing CpG probe arrays for 533 patients with 485,515

probes being shared in common among all of the studies. Methylation data is measured as □-values with 1 demonstrating strong methylation signal and 0 representing strong demethylation. Each dataset additionally contains the sex and age of each patient. I will demonstrate my considerations for potential batch effects, collinearity, linear and non linear interactions, and computation.

Characteristic	n
Sex	
Female	225
Male	308
Study	
Study 1	76
Study 2	106
Study 3	74
Study 4	134
Study 5	143

Figure 1. Study Counts

Methods

The first challenge to address is that the data comes from many different studies. As an initial check, the CpG levels are assessed to determine whether any significant batch effects from different data being sourced from different labs. A subset of the CpGs are taken to screen whether any processing is necessary; as demonstrated by Figure 2, there is no strong concern for needing to take action in this regard.

The next concern is the age distribution of the various studies. Studies 1 and 2 demonstrate a lower age range than studies 3, 4, and 5 (Figure 3). Through PCA, Figure 4 demonstrates that although there is not a clear overlap among all of the studies, there does exist some overlap but with no strong distinction (Figure 4).

Additionally, I later identified through my elastic net model that the effect of study is found to be low.

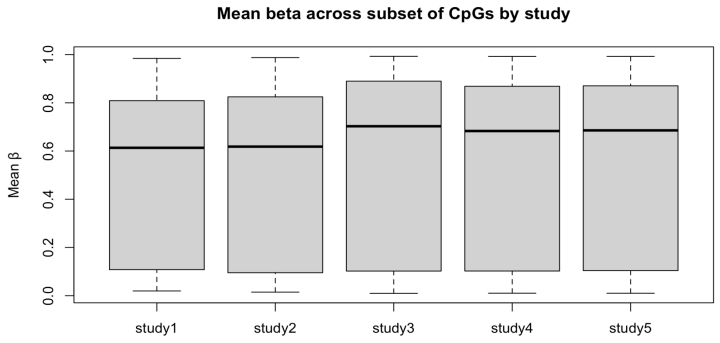


Figure 2. CpG Levels Across Studies

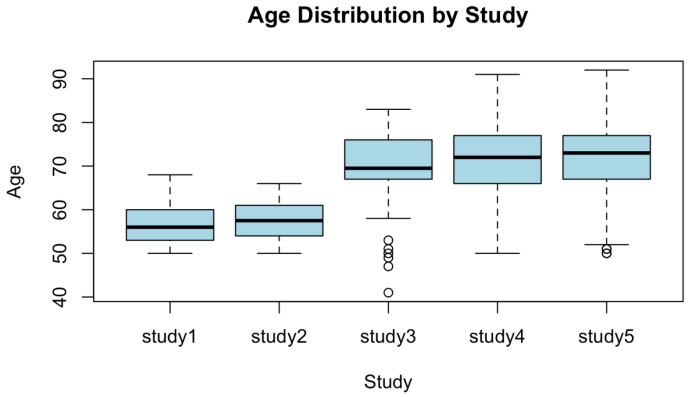
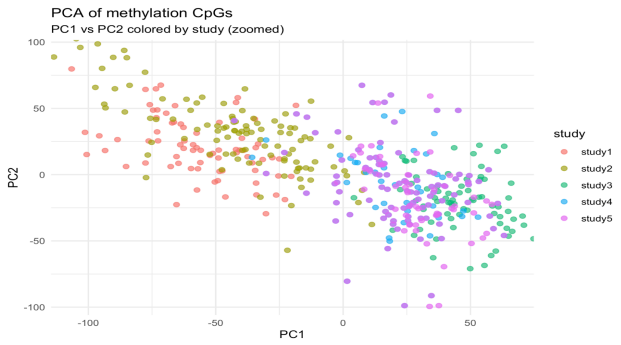


Figure 3. Age Distribution by Study



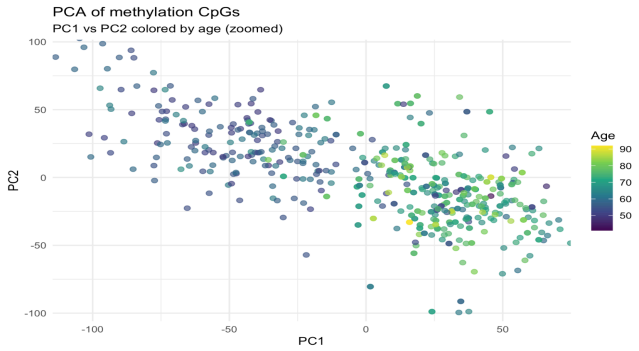


Figure 4. PCA by Study and Age

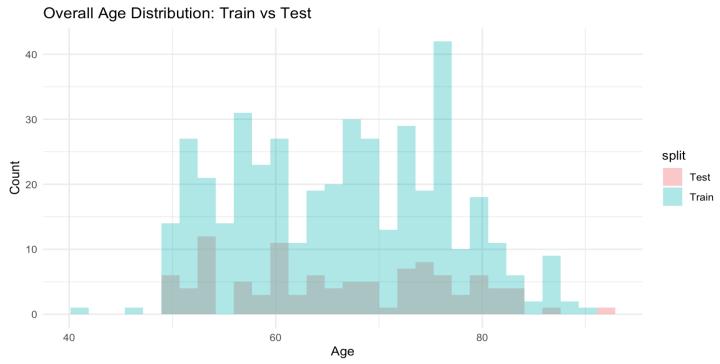
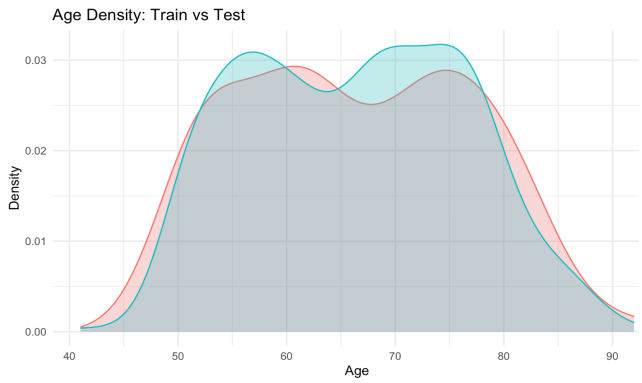


Figure 5. Age Distribution and Counts

Data Processing

After merging the datasets of all five studies, the total number of CpG sites in common is 485,515. I first identified missing values in the data and found around 5,000 columns were affected, and chose to remove them since they represent only a small fraction of the entire set of features. Additionally, none of the dropped features were previously identified by the baseline Horvath clock as significant methylation sites, bolstering my decision to forgo these regions. The datasets were then merged and study labels were added to each row to indicate the respective datasets they came from. The data is then split into a train and test set with an 80:20 split (Figure 5) with the distribution of age being similarly represented in both.



Data Considerations

To begin our exploration, we first make certain considerations and assumptions of the dataset and its biological attributes to acquiesce my initial approach. Given that the data is taken from CpG sites, among the 485,000 genes included, it is highly likely the majority of genes are not associated with age since a significant portion of CpG sites can be attributed to housekeeping genes. Additionally, it would be strange for the majority of genes to be related to age since this would imply we have many more huge drastic changes in our biology throughout our lives than we do in reality.

Our dataset also ranges from age 41 to 92 which does not have nearly as strong developmental transitions as early infancy and childhood development. Therefore, we can first assume that it is not unreasonable to expect a certain degree of sparsity among the features. Finally, as discussed previously, many genes interact with each other. Not only do they interact, but also many come together and comprise common pathways with one another. As a result, we can expect a certain level of collinearity among the CpG sites.

Model 1

Considering all of these aspects, a suitable approach for our first model is to employ elastic net. Elastic net allows us to balance the benefits of lasso to select sparse features by shrinking coefficients of unuseful ones to zero, while also having the capability of ridge to be able to identify sets of features exhibiting collinearity. Additionally, the regularization parameter and alpha parameter of elastic net can be chosen or trained to fit our needs.

Before feeding the data to the elastic net, I first perform filtering processes on the data from the training set. I first assess the variance of the dataset (Figure 6) and observe various levels of filtering. In order to only remove only data that provides insignificant contribution to the model, I choose to filter out 20% of the lowest CpGs with respect to variance. I then perform an additional filter based on correlation and consider multiple levels as well, ultimately deciding to use the filter to target 100,000 genes in the final model based on the results of different thresholds.

The processed data is then fed to my elastic net where sex is included as a covariate, and study ID is encoded through effect coding where deviations from the baseline sum to zero for the studies. I then train the elastic net with standardization among columns using 10-fold cross validation to obtain my regularization term. I also attempt two separate approaches where in one I fix alpha to be 0.5 in order to have a defined balance between the benefits of lasso and ridge, and another one where I allow the model to determine the alpha it deems most optimal.

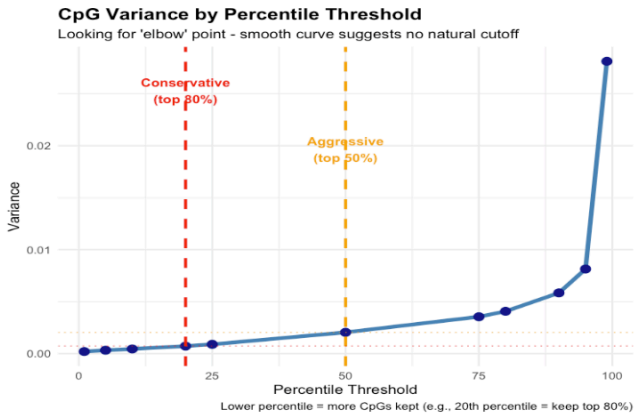


Figure 6. CpG Variance

Model 2

The second model I attempt utilizes boosted trees through XGBoost where in addition to sequential learning on the residuals, XGBoost also introduces a regularization term to reduce bias. I attempted two approaches for feature selection/dimensionality reduction prior to feeding the data to the gradient boosted trees. The first approach I tried was to feed features selected by elastic net into XGBoost. Unfortunately however, this would eliminate many potential nonlinear relationships the features may have with the target and only leave us with linear relationships identified by the model that may or may not still have nonlinear patterns among the survivors. As a result, I finalize the model to instead use PCA as our method of dimensionality reduction on the training set.

PCA is a reliable choice for weeding out uninformative parts of the data; however, its weakness is that unlike elastic net it is not cognizant of the target variable and only selects parts of the data based on the overall contribution to variance of the system. In this particular conjunction with XGBoost though, its

strength is that it does not force us to lose nonlinear relationships with the CpG sites. I applied PCA on the training set to retain the top principal components explaining 90% of the variance. The data was then fed to the XGBoost model where we keep sex as a dummy variable, but do not include study since we do not have very strong reason to believe a significant impact of the relationship between study and age will exist within the model as demonstrated not only in the previously mentioned PCA visualizations from Figure 3 but also in the results of our elastic net. The XGBoost is then trained via 5-fold cross validation with hyperparameter tuning defined by a custom grid.

Model 3

The final model I propose utilizes random forest and considers it as another approach to capture nonlinear interactions. Although elastic net shares the same weaknesses as a feature selection tool as with XGBoost, there is still something to be said about using it for random forest. Elastic net is a useful tool when among many features, sparsity is expected where few features have strong signal. On the other hand, random forest is suitable for settings where many of the features are useful indicators.

Additionally, although elastic net loses many nonlinear relationships, as demonstrated by AltumAge, there is a potential to pick up nonlinear relationships among the features that were selected through only linear ones. AltumAge demonstrated this in the case where among the CpG sites selected by the Horvath clock, some did in fact have nonlinear interactions that could be exploited. The caveat, however, is there may not be much extra leftover signal to extract from nonlinear interactions. I ultimately feed the elastic net feature selected data into the random forest and use 5-fold cross validation with hyperparameter tuning. A custom defined grid of hyperparameters was used to train the final model.

Results

Filtering

Model performance trends regarding how relaxed or aggressive the variance and correlation filtering steps are allowed to be produced expected behavior in the accuracy of the respective levels (Figure 7). When the variance filter and correlation filter are allowed too liberally, the model inevitably loses some important features as shown by an improvement in MAE when the variance filter is lessened but correlation filtering is kept the same. The data further demonstrates that as we keep the variance filter the same but further relax the correlation filter, we sacrifice some accuracy from MAE but gain more precise estimates as shown by the improvement of RMSE demonstrating a reduction in large errors.

Finally, it is shown that by being too relaxed on our filters, the test accuracy decreases but the train accuracy increases, suggesting the model begins to suffer from overfitting that is illustrative of noisy estimates expected from high dimensional problems where p>>n. Interestingly enough, the model selected 3 CpGs present in the chosen CpG sites identified by Horvath's clock at a variance filter level of 80% and correlation filter of top 100,000, but had no increase in identified genes when the filter was relaxed which further suggests against being too

liberal in allowing too much noise into the model.

Model	Variance filter	Correlation filter	Test corr. <dbl>	Test MAE <dbl>	Test RMSE <dbl>	Train MAE <dbl>
Aggressive	Top 50%	Top 20k	0.939	2.73	3.64	1.09
80%–20k	Top 80%	Top 20k	0.947	2.51	3.43	1.09
Conservative	Top 80%	Top 100k	0.951	2.53	3.32	0.82
150k	Top 90%	Top 150k	0.943	2.75	3.57	0.75

Figure 7. Change in MAE Based on Filtering Levels

Metric	Horvath's Clock
Correlation	0.716
MAE	6.36 years
RMSE	7.85 years

Figure 8. Horvath's Clock Baseline

Model	Test MAE (years)
Elastic Net	2.53
XGBoost	2.63
Random Forest	3.24

Table: *Test-set mean absolute error (MAE) by model.*

Figure 9. Test MAE

Models

Our first model using elastic net produced the strongest results with the lowest RMSE and MAE among the three models. Additionally, although study was included through effect coding in the model, it was identified to only have a very small impact on the predictions. Also, comparing the elastic net models where α is fixed at 0.5 in contrast to when α is allowed to be tuned by the model only demonstrated an insignificant increase in test MAE and RMSE. Therefore, in the final model used against the unknown dataset, I purport that, for the purposes of our task, setting α=0.5 is more reliable to have equal balance of the benefits of both ridge and lasso rather than rely on an α that was tuned to perform well on only the training data.

The XGBoost model fed from PCA actually had comparable performance to elastic net and at times slightly better results with certain levels of hyperparameter tuning; however, it had the worst overfitting which signals that the boosting procedure could be dangerously focusing on too much noise and may have issues with generalizability.

Lastly, the random forest fed from elastic net had disappointing performance, ranking the lowest among the three in terms of accuracy; however, the accuracy itself wasn't particularly bad (Figure 9). This most likely means that as we suspected, there were not enough leftover nonlinear interactions left after the feature selection process using linear relationships in elastic net.

Discussion

Although our model was able to achieve satisfactory performance on the dataset after dropping any columns with NA's, it is worth considering that some of these columns may in fact be significant contributors to predicting age. Future projects could try to impute the means of these columns or other methods that account for missing data. Another limitation is that the use of variance filtering in some of the models and the use of PCA for dimensionality reduction do not take into account the target variable, and so may indiscriminately remove CpG sites that in fact would be beneficial to keep. In other

projects, data annotation of the genes can be used to identify not only the gene itself but also the pathways that gene is involved in which could allow for more specialized approaches to the methylation clock modeling process.

It may also be important to note that although study was determined to not be significant in the elastic net model, effect coding does somewhat bias our model towards older estimates on unknown datasets where the model would have to assume a reference of 0 since datasets 3, 4, and 5 tend to have older individuals than studies 1 and 2.

Each model itself also has their own strengths and limitations. Elastic net is able to tackle overfitting with shrinkage through its regularization term and its flexibility between lasso and ridge allows for a mix of feature selection through sparsity and handling correlated features. Additionally, it is interpretable and computationally efficient. On the other hand, it may miss nonlinear relationships and also can miss interactions between genes.

XGBoost is beneficial in its capability to handle regularization and shrinkage as well as capture nonlinear relationships. It can also capture interactions between genes, handles outliers well, and is computationally efficient. It is limited in that the boosting procedure makes it prone to overfitting; dimension reduction methods are necessary to avoid the model from focusing on noise which is of particular concern in high-dimensional problems. It also requires more careful tuning of hyperparameters.

Random forest is another tree based method that also is able to handle nonlinear relationships, can be computationally sped up with parallelization, and capture interactions between genes. It uses bootstrapping sampling to reduce variance, is robust to outliers, and the trees are decorrelated. Similarly to XGBoost, it is important for us to try to use dimension reduction techniques before using random forest. Another weakness as mentioned before is that in the model we used, nonlinear relationships may be largely lost due to using elastic net for feature selection.

Conclusion

Horvath's clock was compared to our models as a baseline, and although my proposed models outperform his on this particular set of data, it must also be noted that his patient population ranged from newborns to people in their 90's and captured the entire range of human lifespan. Therefore, his genes take into account ages not considered in our data that are important in early development for example, but subsequently have less relevance to our age group.

Ultimately among the three models I finalized for the purposes of the predictions on our unknown dataset, elastic net behaves the most reliably and accurately. XGBoost provided the next strongest results on the test set; however, its generalizability may be put into question due to the overfitting that was apparent between the training and testing sets. The random forest also had strong results but demonstrated weaker performance than XGBoost. Additionally, its accuracy was worse than solely using elastic net, but had less issues with overfitting than XGBoost. Overall, elastic net appears to be the most reliable model, followed by random forest, and lastly XGBoost due to overfitting.

References

1. Insights to aging prediction with AI based epigenetic clocks - PMC
2. Universal DNA methylation age across mammalian tissues | Nature Aging
3. Integrating Epigenetic and Phenotypic Features for Biological Age Estimation in Cancer Patients via Multimodal Learning
4. DNA methylation-based biomarkers and the epigenetic clock theory of ageing | Nature Reviews Genetics
5. DNA methylation age of human tissues and cell types | Genome Biology
6. Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex - PMC
7. Why do tree-based models still outperform deep learning on typical tabular data?
8. Benefits of dimension reduction in penalized regression methods for high-dimensional grouped data: a case study in low sample size | Bioinformatics | Oxford Academic
9. Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction | Journal of Animal Science and Biotechnology
10. EWASex: an efficient R-package to predict sex in epigenome-wide association studies | Bioinformatics | Oxford Academic
11. Cross Hybridization - an overview | ScienceDirect Topics