Finding Biomarkers of Biological Age

2024-04-30

Introduction

Background

Biological Age can be thought of as a measure of how well an individual's physiological systems are functioning. In a way, it is the "age of your cells" as opposed to your chronological age from your date of birth. Biological age may be a better predictor of health than chronological age and so the ability to quantify biological age would allow for more effective implementation of age-related interventions and better prediction of age-related conditions. As such, we wish to identify potential biomarkers in the blood that could act as predictors of biological age.

Data

The dataset we are using is titled "Health Test by Blood" and contains cardiovascular and kidney function information from 5132 patients. There are 11 variables in the dataset including:

- Age: age in years of participants
- Gender: M or F
- BMI
- Chol: Cholesterol levels (mmol/L)
- TG: Triglyceride levels (mmol/L). Triglycerides are a type of fat found in the blood. High levels can increase the risk of heart disease.
- HDL: High-density lipoproteins (mmol/L). This is the "good" cholesterol in your blood. High levels indicate good heart health.
- LDL: Low-density lipoproteins (mmol/L). This is the "bad cholesterol in your blood.
- Cr: Creatinine (mmol/L). Creatinine is a waste product from muscle metabolism and a measure of kidney function.
- BUN: Blood urea nitrogen (mmol/L). A measure of urea nitrogen levels in the blood that is an indicator of kidney and liver function.
- Diagnosis: Type II Diabetes Diagnosis (1:Yes or 0:No)

There was no missing data in this dataset. There was 1 observation where the gender was marked "f" instead of the capital "F". We believe this was a data entry error and recode this as "F".

Study Aims

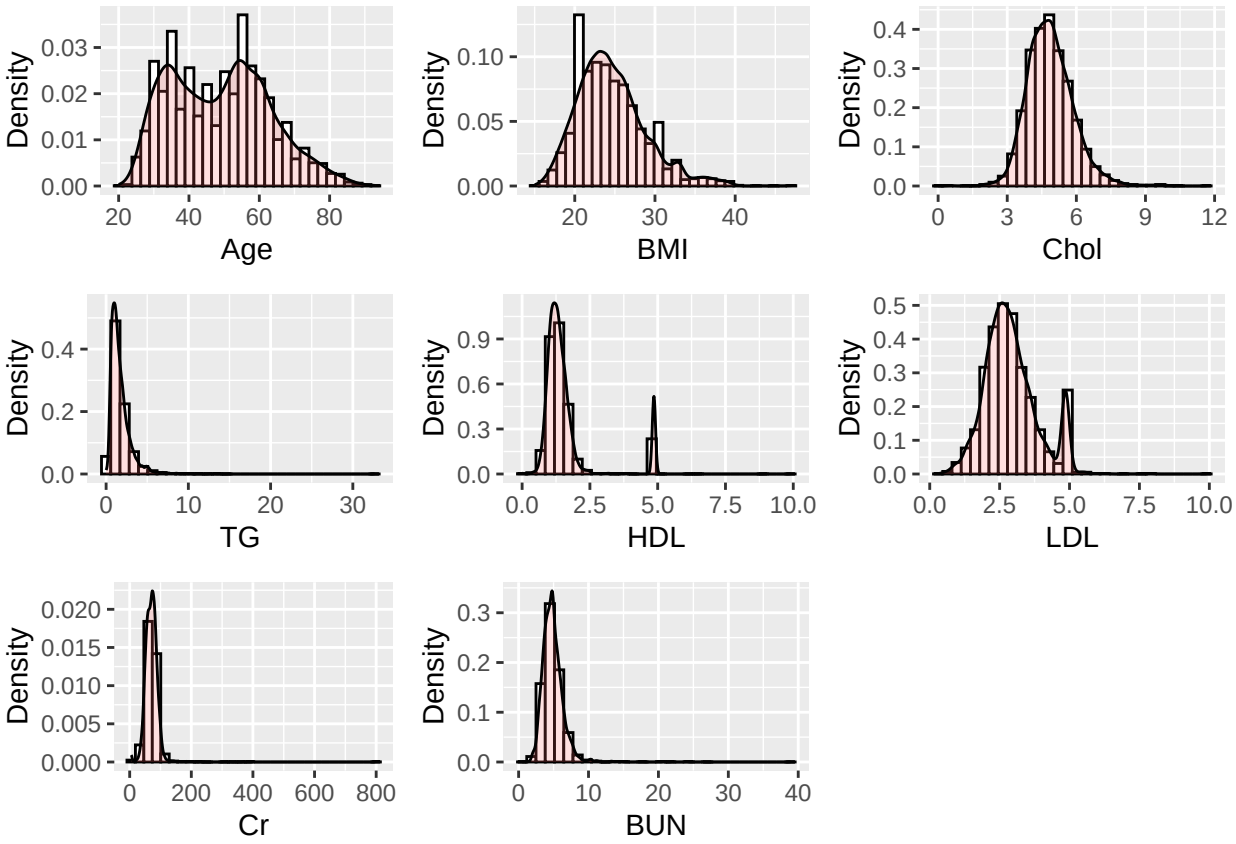
While our primary goal was to find biomarkers to predict biological age we do not have a measure of this available in our dataset. As a result, we use chronological age in our model development and training. We wish to first build up this framework for quantifying feature importance using Lasso regression and machine learning methods to identify some potential blood biomarkers as a starting point. When there is biological age data available our framework can be easily applied to this new data, refined, and the biomarkers can be compared.

Table 1: Summary Statistics

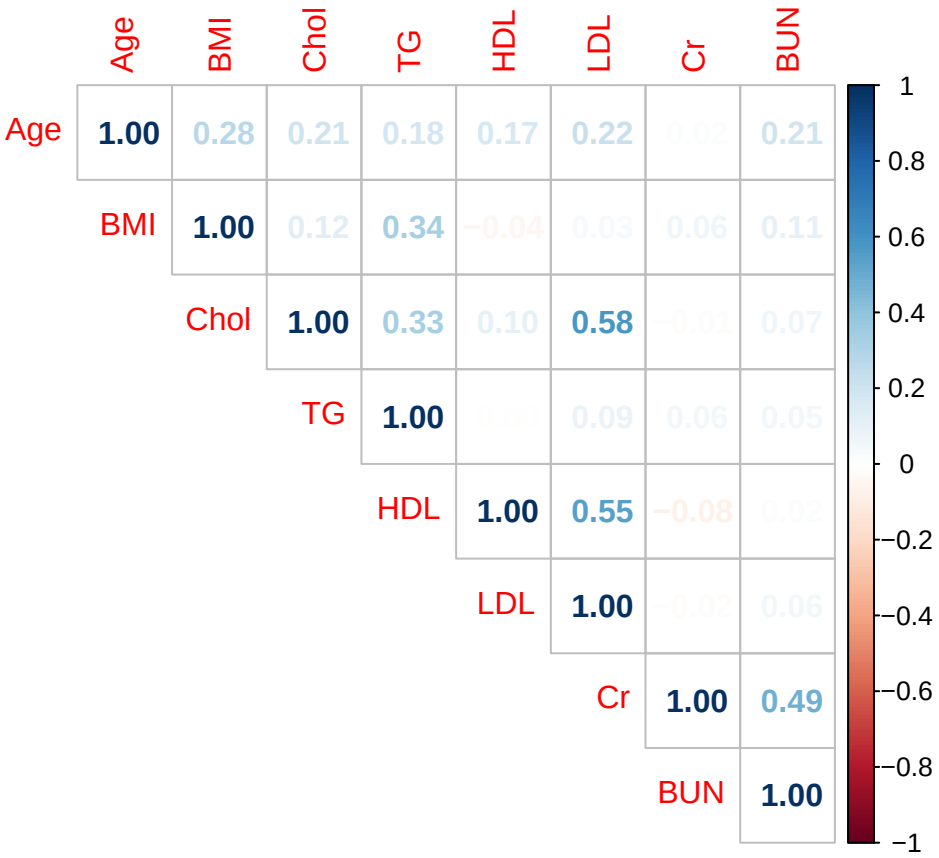
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	5132	49	14	20	36	59	93
Gender	5132						
... F	1876	37%					
... M	3256	63%					
BMI	5132	25	4.3	15	22	27	47
Chol	5132	4.9	1	0	4.2	5.5	12
TG	5132	1.7	1.3	0	0.91	2.1	33
HDL	5132	1.6	1	0	1.1	1.6	9.9
LDL	5132	2.9	0.95	0.3	2.3	3.4	9.9
Cr	5132	71	28	4.9	58	82	800
BUN	5132	4.9	1.7	0.5	3.9	5.6	39
Diagnosis	5132						
... No	3139	61%					
... Yes	1993	39%					
AgeCat	5132						
... [20,35)	1027	20%					
... [35,50)	1551	30%					
... [50,65)	1844	36%					
... [65,80)	618	12%					
... [80,95)	92	2%					

Exploratory Data Analysis

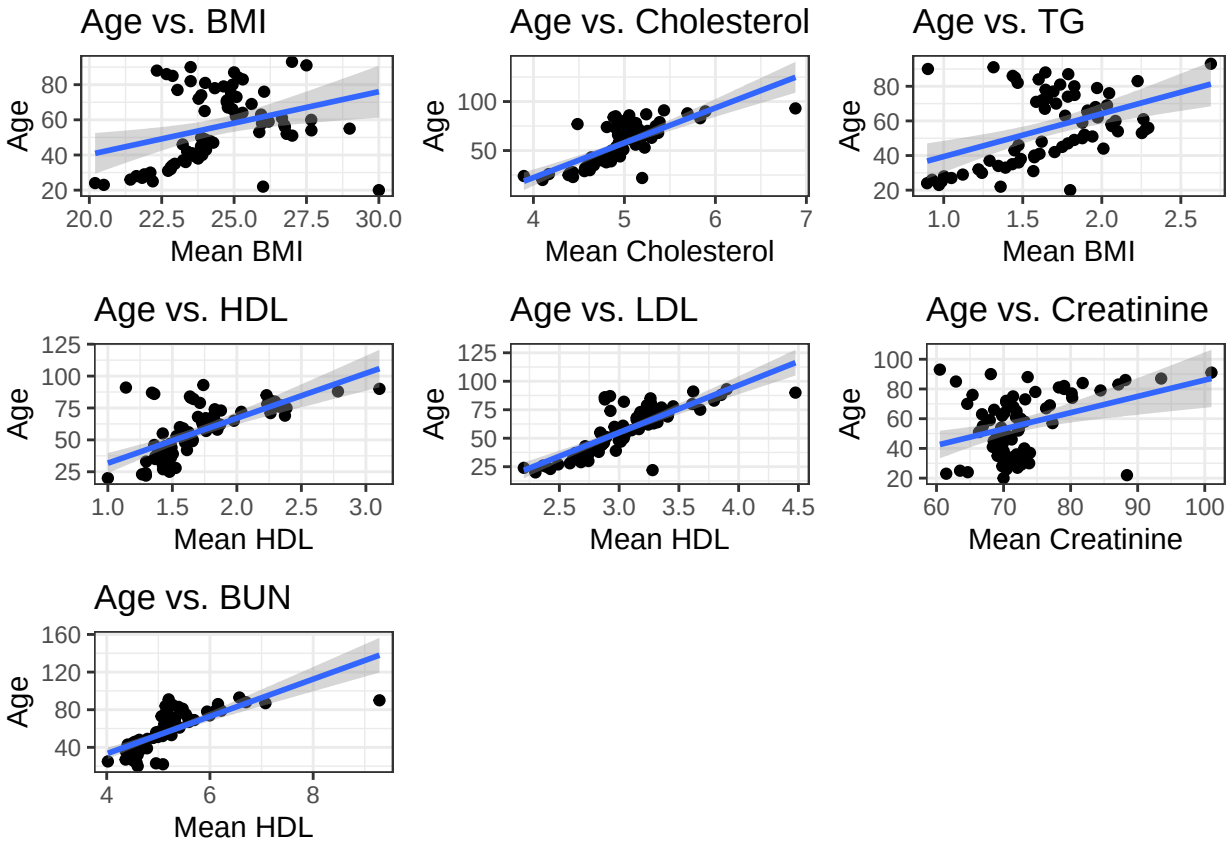
Above we have a table with summary statistics of all variables in our dataset. We created an age category variable spanning ages 20-95 with 5 categories of length 15 years each.



Above we have plots of the densities of each continuous variable. We see that most variables appear to be fairly normally distributed with some variables like Cr and TG having a few very large entries. Upon investigation these entries are not outside the realm of possible values and thus we keep them in our dataset.

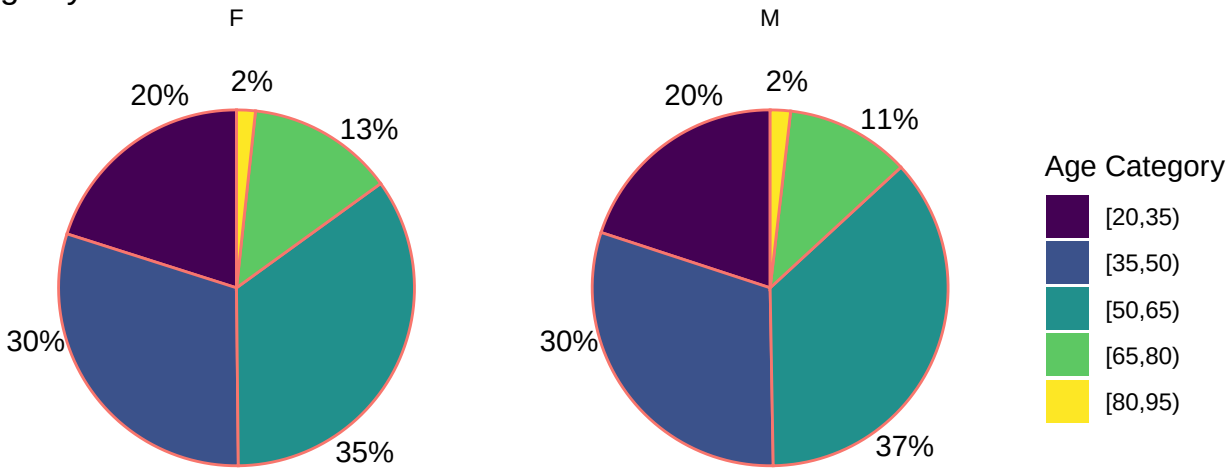


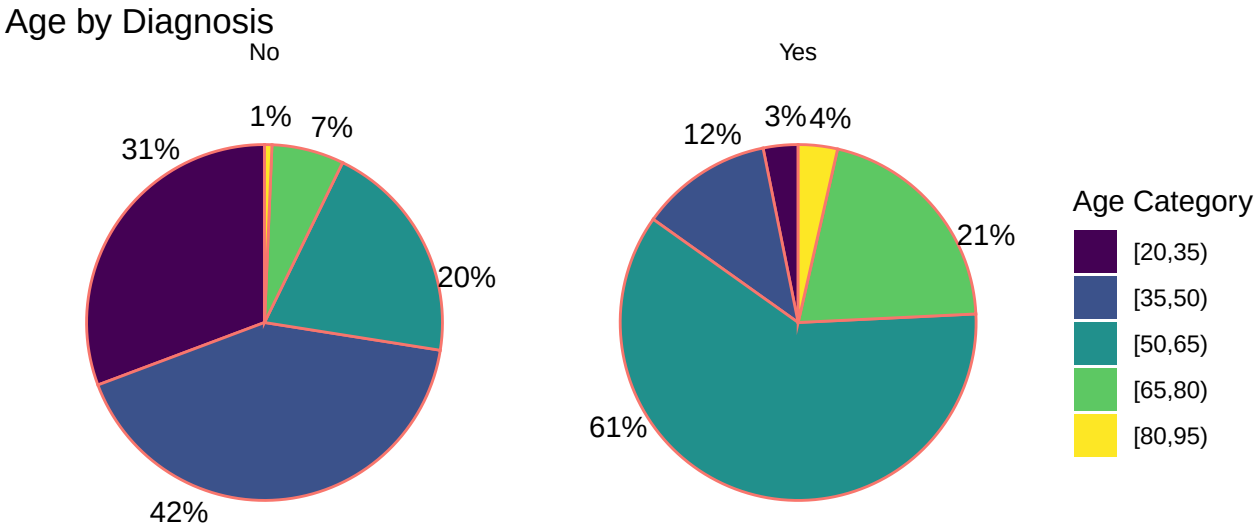
Above is a correlation plot of the variables in our dataset. Generally, there does not seem to be much correlation among the different values. We note that LDL is mildly correlated with both Cholesterol and HDL, but HDL does not appear to be correlated with Cholesterol.



Now above we have plots age against mean values of each of our variables. All variables appear to have a positive association with age with most notably Cholesterol, Triglycerides, and BUN having the greatest apparent association.

Age by Gender





Above are 2 figures with pie charts. The first figure depicts the age makeup of the female and male participants. The percentages of participants in each age category look very similar across the two gender in the dataset. The second figure is the age makeup of the participants with and without a diabetes diagnosis. Here, we note that participants with diabetes tend to be much older

Methods

In this project, linear regression with L1 penalty and two machine learning methods (Random Forest and Gradient Boosting Machine) are used to analyze the importance of features in predicting age.

Lasso Regression

We first implemented from scratch a linear regression with L1 penalty using coordinate descent and soft-thresholding.

The objective function to minimize for Lasso Regression is: $\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$ where λ is the hyperparameter for L1 penalty, n is the number of observations, p is the number of features, y_i is the response age for the i-th observation, x_{ij} is the value of the j-th feature for the i-th observation, and β_j is the coefficient of the j-th feature. The Iterative Shrinkage Thresholding Algorithm (ISTA) and coordinate descent are used for the estimation procedure, which is essentially updating each coefficient using the proximal gradient descent for the objective function.

Since LASSO can effectively perform feature selection by setting some coefficients to exactly zero, it is very helpful in identifying the most important features for predicting age.

To get the feature importance scores in Lasso Regression analogous to the machine learning methods, we used bootstrap for statistical inference. 500 bootstrap samples were generated from the training data and Lasso Regression is fitted on each sample to obtain the coefficients. The importance score of each feature was calculated as the proportion of times it was nonzero across all bootstrap samples. Features with higher importance scores, i.e. those more frequently selected, are considered more important in the Lasso regression. The lasso_bootstrap_inference function was used to perform the bootstrap inference. It returns the probability of each feature being selected in the Lasso model across the bootstrap samples as described above.

Random Forest (RF)

Random Forest is an ensemble method that combines multiple decision trees to make predictions. The predictions of each tree are averaged to obtain the final prediction. It is a robust machine learning algorithm that is compatible with continuous outcome with both continuous and discrete inputs. It also provides a measure of feature importance based on the decrease in impurity, which could help identify important features for age prediction. Specifically, the decrease in impurity (Gini impurity by default) when splitting on that feature is used to determine the importance of each feature. Features with higher importance scores are considered more important in the Random Forest model. We used the `caret` package in R to train the Random Forest model. The do_rf_v2 function from the package was used to fit the model with the default tuning procedure.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine is another ensemble method that builds an additive model of weak learners, in this case, decision trees. It is also a robust machine learning algorithm providing a measure of feature importance based on the improvement in model performance. Specifically, the improvement in the model's performance attributable to each feature is measured to calculate the feature importance scores in GBM. Features with higher importance scores are considered more important in the GBM model. We used the `gbm` package in R to train the GBM model. The do_gbm_v2 function from the package was used to fit the model.

Comparison of Feature Importance Scores

To compare the feature importance scores across the different models, we ranked the features based on their importance scores from each model dnd then compared the rankings to identify features that are important across the models. It is noteworthy that given the variable selection property of LASSO, the lasso regression feature importance scores have exact zeros, which is rarely the case in machine learning algorithms.

After obtaining the feature importance scores from each model, we identified the top features with the highest scores for each approach. We then refitted the three models using only those top features and used them to predict age in the test set. Model performance was assessed by comparing metrics such as mean squared error (MSE), mean absolute error (MAE), prediction R squared in the training set (R2_training), and prediction R squared in the test set (R2_test) as they are good measures of the models' predictive accuracy and generalization ability.

Results

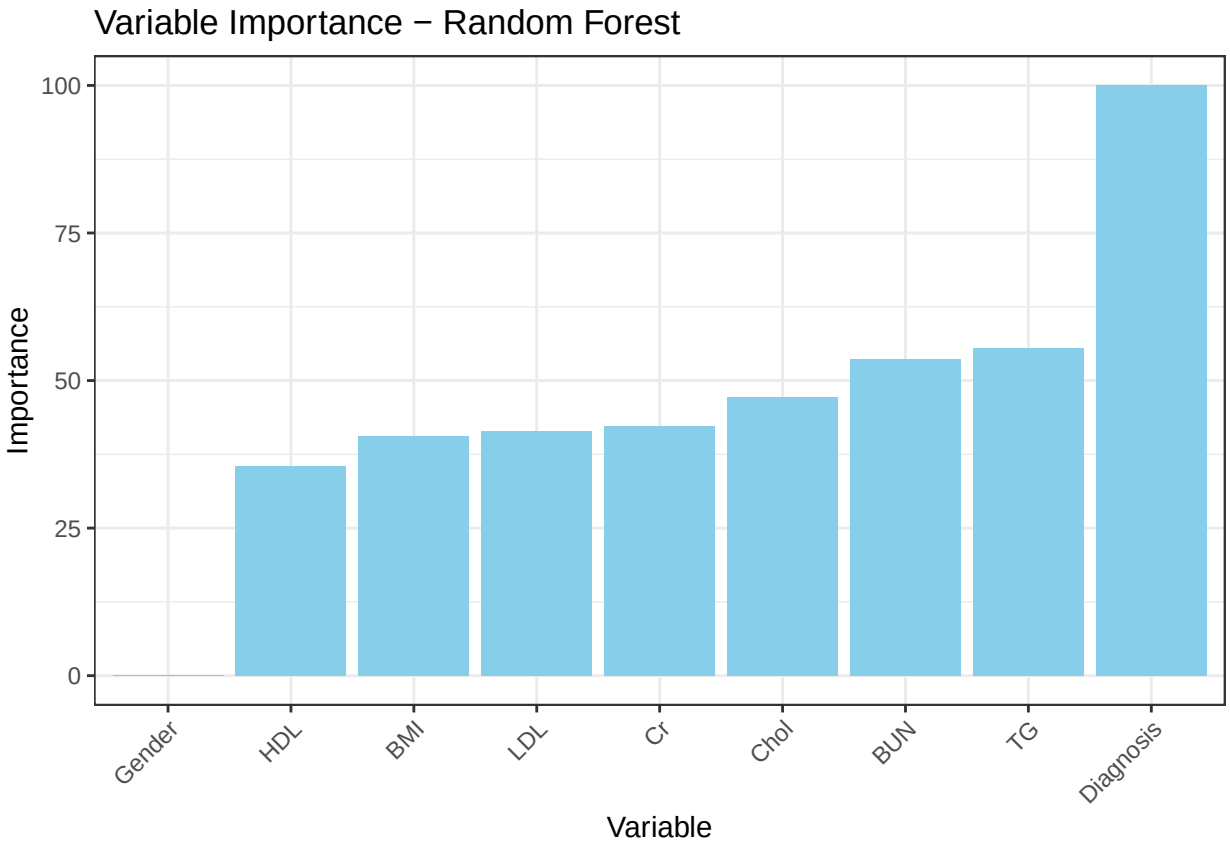
To establish our models, we initially partitioned the complete data into a 75% training set and a 25% test set. This data splitting was executed by creating an index denoting whether the subject being in the training set using `rbinom()` function, with a probability of 1 set to 0.75. Subsequently, we corrected the gender classification for an individual erroneously labeled as "f", which should have been "F" to indicate female gender. Furthermore, we converted both gender and diagnosis variables into factors for the following analysis.

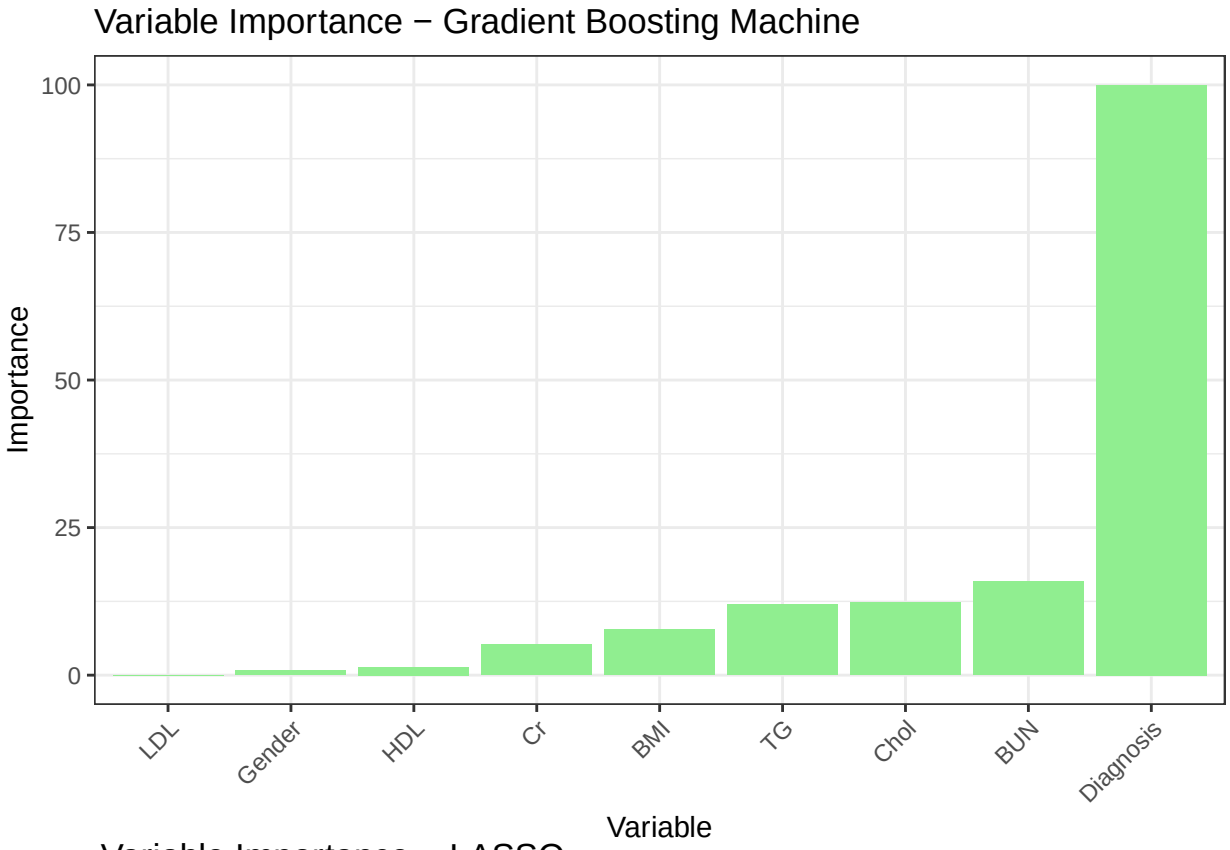
Upon completing data preparation, we proceeded to construct LASSO, GBM, and random forest models, incorporating all nine features. Codes for model fitting procedure are attached in the following chunk. In order to optimize parameters for GBM, several exploratory tests were conducted to pre-determine the suitable tuning range. The outcomes of these preliminary tests were consolidated into `gbm_tg`. For the random forest model, we employed the default tuning procedure provided by `caret::train`. In the case of the LASSO model, feature importance scores were computed utilizing bootstrap inference as detailed in the method section.

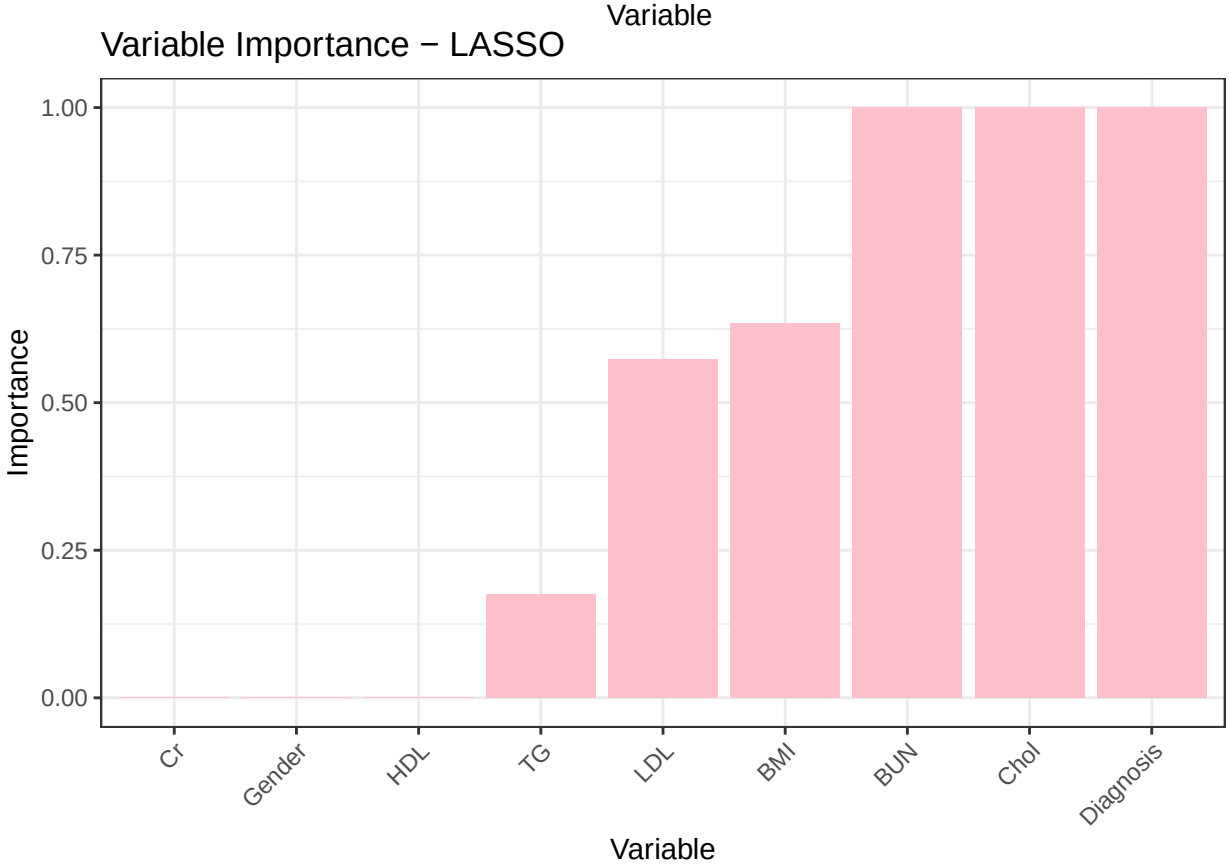
Table 2: Model Variable Importance Scores

Variable	rf	gbm	lasso
BMI	40.507	7.705	0.634
BUN	53.622	15.908	1.000
Chol	47.144	12.322	1.000
Cr	42.215	5.155	0.000
Diagnosis	100.000	100.000	1.000
Gender	0.000	0.798	0.000
HDL	35.500	1.384	0.000
LDL	41.506	0.000	0.574
TG	55.423	12.003	0.176

Once training procedure finished, feature importance scores from the three models were obtained. Here, we summary the scores into the following table and figures:







Based on the feature importance scores provided by each model, we identified the top 5 features with the

Table 3: Model Performance Comparison

Model	MSE	MAE	R2_training	R2_test
LASSO	136.453	9.226	0.308	0.332
RandomForest	138.319	9.193	0.305	0.340
GBM	128.117	8.888	0.345	0.371

highest scores for each approach.

The top 5 features for LASSO model are diabetes diagnosis, cholesterol level, blood urea nitrogen, BMI, low-density lipoproteins levels. For GBM model, these features include diabetes diagnosis, blood urea nitrogen, cholesterol level, triglycerides level, and BMI. For random forest, the top 5 features are diabetes diagnosis, triglycerides level, blood urea nitrogen, cholesterol level, and creatinine level. Notably, diabetes diagnosis, cholesterol level, and blood urea nitrogen consistently appear among the top 5 feature lists across all models, indicating robustness across different methodologies.

Next, we refit the three model using only the top 5 features and employed the newly trained models to predict age in the test set. Model performance were assessed by comparing metrics such as mean squared error (MSE), mean absolute error (MAE), prediction R squared in the training set (R2_training), as well as prediction R squared in the test set (R2_test). The results summarizing the model performances are presented in the subsequent table.

Overall, GBM has best performance compared to the other two models. Random forest and LASSO have comparable performance. Specifically, the LASSO model boasts a smaller MSE and higher R^2 in the training set, whereas the random forest model achieves a smaller MAE and higher R^2 in the test set.

Conclusion

In this study, we utilized three distinct models, namely LASSO, GBM, and random forest, to predict age based on various clinical features. Through feature importance analysis, we identified key predictors for each model and selected the top 5 features with the highest importance scores. Subsequently, we refined the models by training them solely on these top features and evaluated their performance using metrics such as MSE, MAE, and prediction R squared. Ultimately, our findings highlight the overall better performance of the GBM model, with consistent performance levels observed between random forest and LASSO models.

Both machine learning models shows better prediction behavior, however, the distribution of their feature importance scores display a notable imbalance. This might be caused by both approaches tending to evaluate the importance of features at least more than zero. Conversely, the LASSO model, although exhibits a relatively small prediction score, demonstrates a more balanced distribution of feature importance scores. This is attributed to LASSO's tendency to assign zero importance to uninformative variables.

In the future, alternative modeling approaches, such as machine learning models employing SHapley Additive exPlanations (SHAP), and linear models incorporating penalties like Dantzig, MCP, and ENet, could be explored for benchmarking purposes.

Although we managed to obtain a meaningful model, our results could not be used as biomarker as the age in our data set does not accurately represent biological age. Therefore, more appropriate data with individuals' biological age is needed for further justification of out models. Moreover, to enhance the generalizability of our findings, it is imperative to validate and evaluate our framework across different external data sets, rather than relying on only the test set within our study.