**Abstract**

Our main research purpose towards this age prediction task is to indicate that beyond blood biomarkers, data variables that come from chronic disease, self-reported functional limitation and cognition measures may help increase the accuracy when applying machine learning algorithms to predict biological age.

**Data construction**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | HouseholdID | 19 Blood biomarkers features | 14 Chronic disease features | 20 self-reported functional limitation features | 8 cognition features | age |

Our datasets consist data from 2011(sample size = 10027) and 2015(sample size = 9668). Both of these datasets contain anonymous IDs and householdIDs to help identify each data provider, and 61 features from blood biomarkers (19 features: white blood cell in thousands, MCV, blood urea nitrogen … etc.), chronic disease (14 features: ever had high blood pressure, ever had diabetes, ever had cancer … etc.), self-reported functional limitation (20 features: Diff-Dressing, Diff-Bathing, Diff-Eating … etc.) and cognition (8 features: CESD Score, immediate word recall, able to draw assign picture … etc.).

**Methodology**

The general steps to perform machine learning algorithms are:

1. Data cleaning

First of all, we filter our age column values by applying a rule of no less than 40 years old and no greater than 85 years old. Secondly, we drop duplicated samples if there are any. Thirdly, we calculate the missing ratio of each column and decide whether to leave or drop the corresponding columns (we usually drop columns that have a rather high missing ratio). Finally, we check all the data types of each columns to make sure that no mistakes like taking numerical features as categorical features or the opposite.

1. Feature engineering

Feature engineering consists:

* Missing value imputation: normally replace with mean, median, mode.
* Separate feature types into numerical type, categorical type and ordinal type and perform preprocessing techniques such as one-hot encoding , ordinal feature transformation etc.
* Normalization: usually use min-max scalar and standard scalar.
* Low variance: Dumping Columns with low variance.
* Multi-collinearity: take multi-collinearity between variables.

1. Model comparison

The first thing we need to be clear is that this age prediction is a regression problem in the machine learning fields. The reason why we do not use deep learning method is due to our insufficient sample size of our dataset.

To indicate the importance of non-blood biomarker features of prediction, we first train models with only blood biomarker features to get the R-Square score and the MSE (Mean Square Error) and then we train models with blood biomarker features together with other non-blood biomarker features. Thus, we can compare the score changes of R-Square and MSE before and after we take non blood biomarker features into consideration.

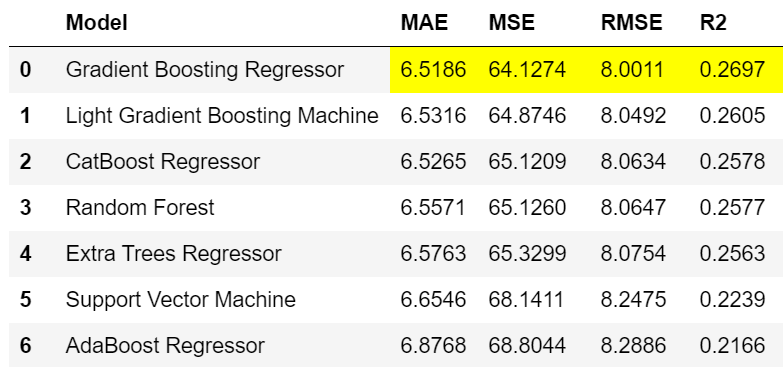
1. Result

We compare different baselines of our selected machine learning algorithms with 10 cross-validation and list seven best models’ performances from high to low:

* Only blood biomarkers:

Dataset construction:

|  |  |  |  |
| --- | --- | --- | --- |
| ID | HouseholdID | 19 Blood biomarkers features | age |

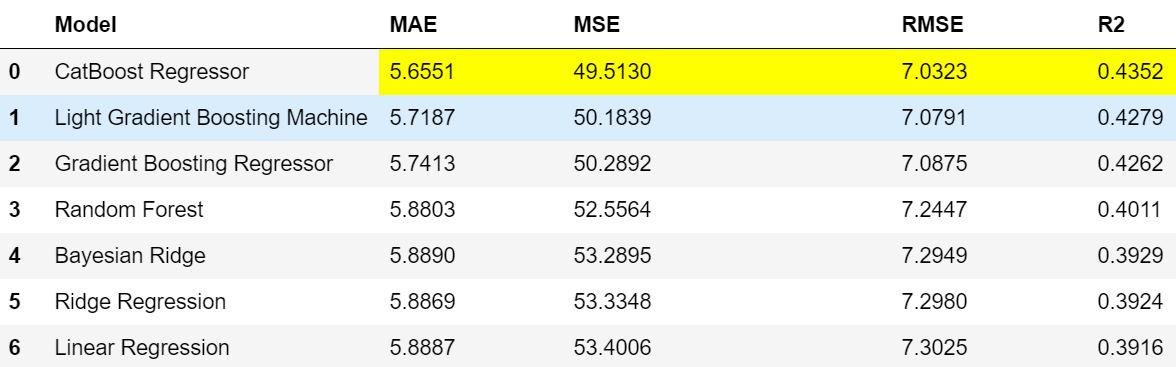


Best regressor is Gradient Boosting Regressor with R-Square = 0.2697, MAE = 6.5186

* After adding non-blood biomarkers:

Dataset construction:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | HouseholdID | 19 Blood biomarkers features | 14 Chronic disease features | 20 self-reported functional limitation features | 8 cognition features | age |

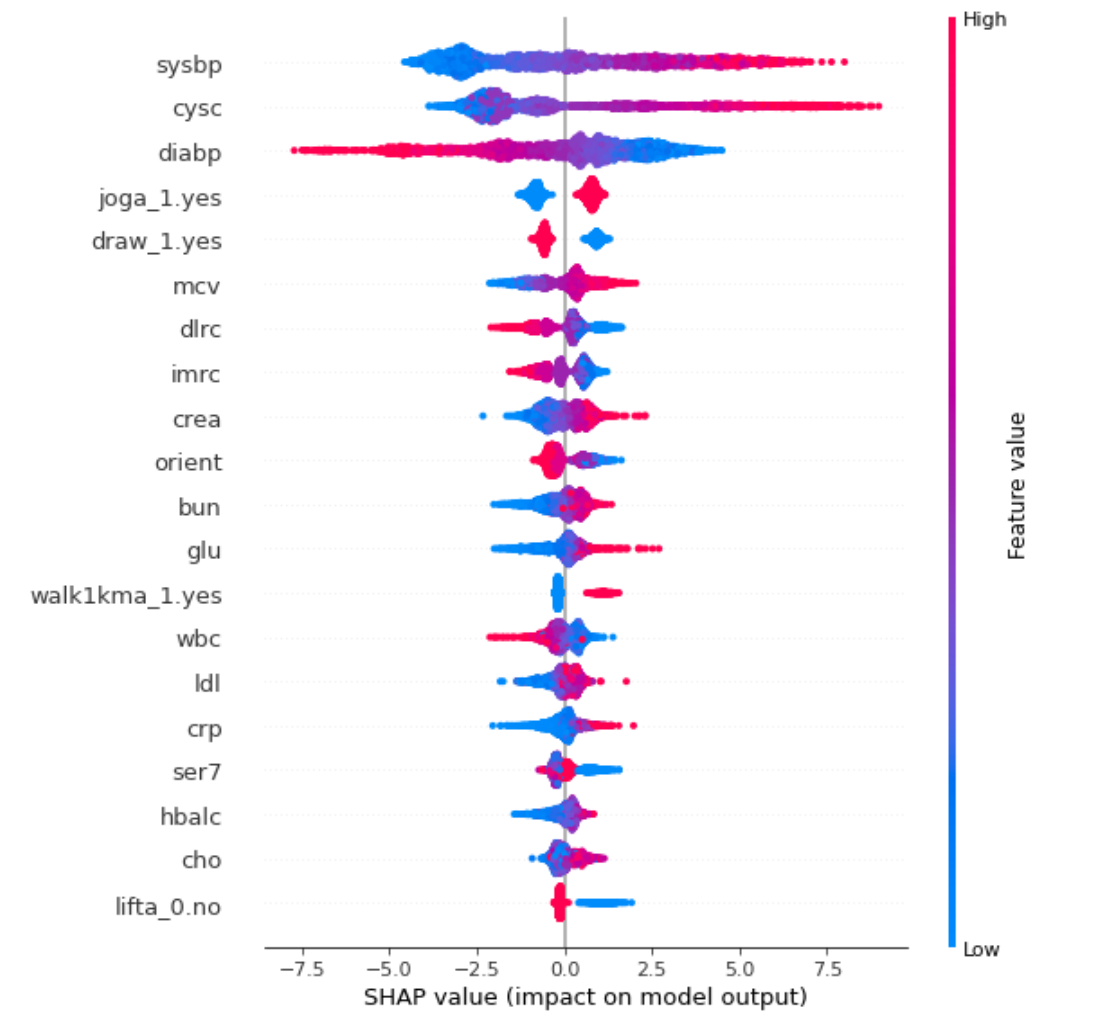


Best regressor is CatBoost Regressor with R-Square = 0.4352, MAE = 5.6551.

**Feature impacts on age**

We implement SHAP value analysis to see the impacts of features on predicting our target variable age. SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. The interpretation of the Shapley value is: Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.

The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The color represents the value of the feature from low to high. Overlapping points are jittered in y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance.

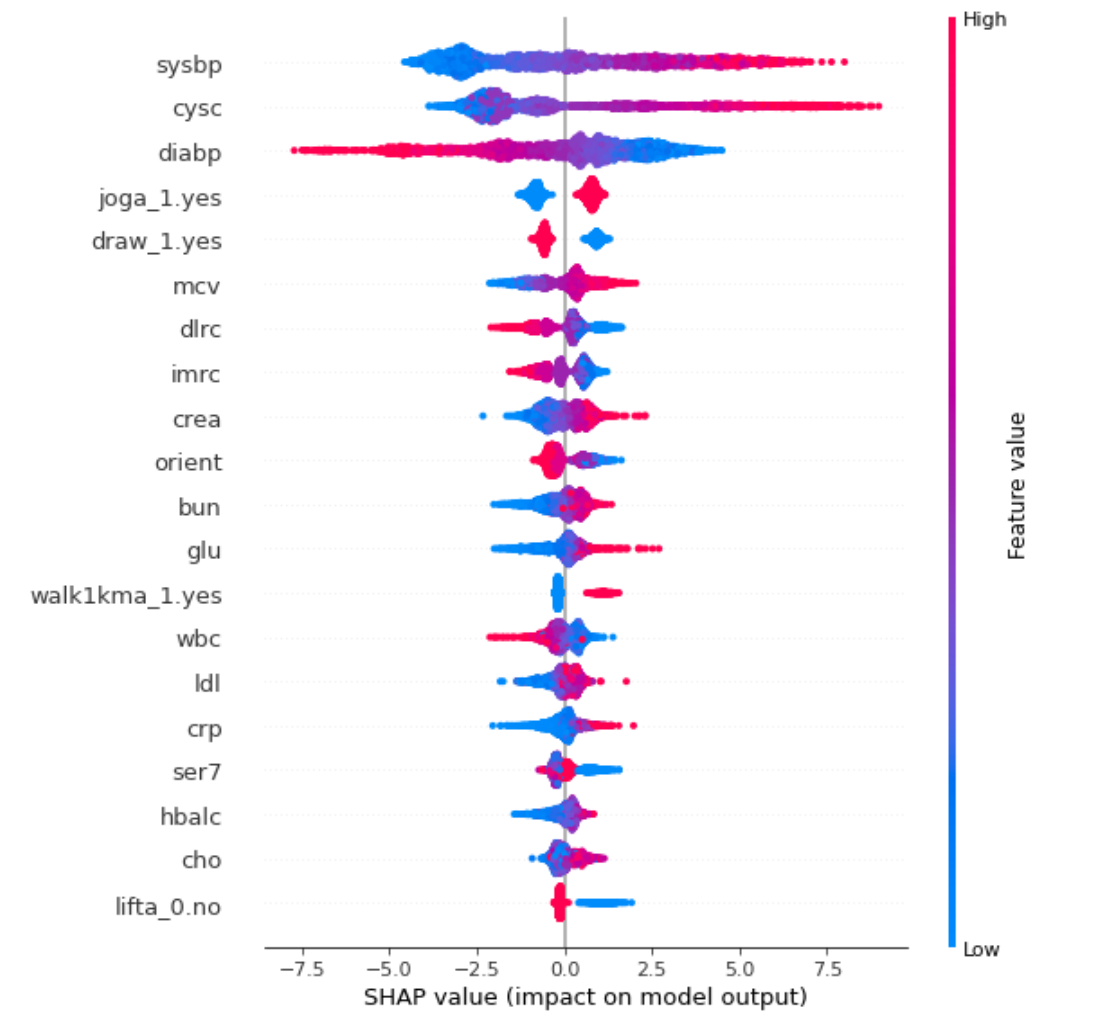


We can see that the top 5 most important features from high to low are:

sysbp, cysc,diap,joga

**Conclusion**

We notice that after adding chronic disease features, self-reported functional limitation features and cognition features, R-square increases hugely from 0.27 to 0.44 and MAE decreases from 6.52 years to 5.66 years, which indicates that non blood biomarker features can increase the prediction ability during the model training process and thus shall be considered as important indicators during age assessment process in the future.

****