CS 760: Machine Learning Spring 2022

Final Project

AUTHORS: Ayushi Mishra, Neil Bhutada

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

Opportunistic Cardiometabolic Screening is a process of collecting Computerized Tomography (CT) data for a pateint. CT scan collects a series of X-ray images for different parts of the body assisted in analyzing Bone, Fat, and Muscle measures, along with Aortic clarification and Liver fat of pateints (Mayo Foundation, p.1). Combining this CT data along with clinical data, such as BMI, Framingham Risk Score (FRS), Metabolic Syndrome etc. of several pateints helped us in predicting in the *biological age* and *number of days before death* of pateints. To compute the Biological age of a pateint we compared their CT and clinical data with other normal (or healthy) pateints within various age groups. For our analysis, we utilized two Machine learning models including Gradient Boosting Machine (GBM) and Weighted K-Nearest Neighbour (Weighted K-NN) to predict *biological age* and *number of days before death* of pateints. After our implementation, we came to a conclusion that incorporating clinical data along with CT data aided in improving our predictions for *biological age* by $\sim 40\%$ and *number of days before death* by $\sim 16\%$. Comparing the two models, we realized that Gradient Boosting Machine (GBM) performed better compared to Weighted K-Nearest Neighbour in our analysis to predict the *biological age* and *number of days before death* with and without clinical data by $\sim 33\%$.

1 Introduction

During Opportunistic Cardiometabolic Screening, alot of the data goes unused after a preliminary analysis. Thus, in attempt to make the Opportunistic Cardiometabolic Screening data obtained more useful, we analyzed the CT scan data to predict the *biological age* and *number of days before death* of a patient. To predict the *number of days before death* we used the patient's calculated *biological age*. The way we predicted biological age was by comparing CT and clinical data of a pateint to other healthy/normal patients of various age groups. For example, a 50 year-old person may have a body (in our case CT and clinical data) similar to other 80 year-olds.

Initially, to compute a patient's *biological age* we filtered out the data to select healthy and alive patients based on various age groups solely using their CT data. The filtered data of healthly patients with their *chronological age* was used as the training dataset to predict the biological age. After creating a model to predict a patient's *biological age*, we used this model to predict the *biological age* of all the dead people in the dataset. Then with the CT data scans and *biological age* of all the dead people, we developed a model to predict the *number of days before death*. Then we incorporated clinical data (such as BMI, Metabolic Syndrome, Tobacco consumption, etc.) about patients to further filter healthy and alive people to predict the *biological age*. Similarly, we created another model to predict the *number of days before death* that uses a patient's clinical data, CT data, and newly calculated *biological age*.

2 Related/Similar Work

Automated CT biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: a retrospective cohort study' by Pickhardt et. al potentially showed how clinical data along with CT data can improve insights and a model's performance to predict clinical outcomes.

3 Dataset Description

The *Opportunistic Cardiometabolic Screening* is owned by Perry Pickhardt (Department of Radiology, UW-Madison). The dataset consists of 9223 records, such that each records in the dataset represents a unique pateint. In our report, we will be using the following features of the dataset for our analysis:

Clinical Data

Body Mass Index (BMI) that determines whether a person has a healthy weight, **Body Mass Index greater than 30 (BMI >30)** shows if a person has a BMI greater than 30 and no means that a person has lower than 30 BMI, **Age at Computerized Tomography (Age at CT)** consists of chronological age reported during the CT scan, **Tobacco** tells if a pateint consumes tobacco or not, **Framingham Risk Score (FRS 10-year risk (%))** is used to estimate 10-yr risk of developing coronary heart disease, **Fracture Risk Assessment Score (FRAX 10y Fx Prob (Orange-w/ DXA))** shows a 10yr probability of fracture at the femoral neck or spine, **Fracture Risk Assessment Score for hip (FRAX 10y Hip Fx Prob (Orange-w/ DXA))** shows a 10yr probability of hip fracture, **Metabolic Syndrome (Met Sx)** tells if a patient has the syndrome or not.

Computerized Tomography Data

Bone measure/BMD (L1 HU) measures the amount of bone mineral density, **TAT Area (cm2)** Total Adipose tissue, **VAT Area (cm2)** Visceral Adipose tissue, **SAT Area (cm2)** Subcutaneous Adipose tissue, **VAT/SAT Ratio** analyzes cardiometabolic risk, **Total Body Area EA (cm2)** of a pateint, **Muscle HU** tells muscle measure in Hounsfeild units, **Muscle Area (cm2)** muscle area measure in cm2, **Skeletal Muscle Index (L3 SMI (cm2/m2))** indicator of muscle depletion, **Aortic Calcification (AoCa Agatston)** measures calcium deposists on the aortic valve, **Liver Fat (Liver HU (Median))** measures extra fat stored in the liver in Hounsfeild units

Clinical Outcomes

Death (DEATH [d from CT]) consists of the number of days after the CT scan when the death of the pateint tool place.

4 Approach

4.1 Preliminary Analysis

We began our analysis by first looking at CT scan data to understand how these features are correlated with age at CT or number of days before death. We saw that none of the features among the CT scan data seems to be correlated with number of days before death. Moreover, we saw that age at CT is slightly correlated with Bone Mineral Density (L1 BMD HU).

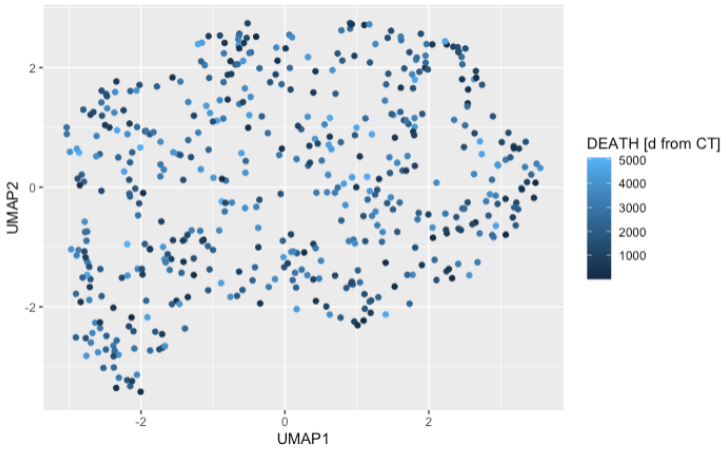

Figure 1: UMAP of CT data

Since there is very little correlation, to further understand the data we performed UMAP - a non-linear dimensionality reduction method - to see if there were any significant clusters in the data. From a very careful observation of the plot in Figure 1 we can see that many points nearby have similar *number of days before death*.

4.2 Making predictions using only CT data

4.2.1 Pre-processing data to predict biological age

To filter out all the people who were healthy and alive, we assumed that all the outliers in the dataset were considered to be unhealthy. To proceed with our analysis, we performed Principal Component Analysis (PCA) on the above mentioned features of CT scan data to remove outliers. Thus, we filtered out the outliers from all the all principal components by only taking data that fell within $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$, where Q1 represents 25th percentile value, Q3 represents 75th percentile value. After performing PCA, we realized that Aortic calcification and Bone Mineral Density (L1 BMD HU) still had some outliers. So, we again captured the data that fell within $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$ for these features.

4.2.2 Model description for predicting biological age

To predict biological age, we used two models - Weighted-KNN and Gradient Boosting Machine (GBM)

Based on our definition of biological age defined above, our model has to look for patients with similar data (CT data scans, clinical data) and calculate their biological age. Furthermore, the model should give more importance to neighbouring data points. Hence, for predicting biological age we used Weighted-KNN as one option.

We implemented Weighted-KNN using the "train.kknn" function in the "kknn" library. The "train.kknn" performs leave one out cross-validation and finds an optimal value of "k" if "kmax" is specified. From Figure 2, the optimal value of "k" seemed to be around 15. Weights are assigned to different neighbours based on the distance from the new observation and the kernel passed. The kernel here means the probability density function. According to the pseudo code, the weights are first normalized and are between $[10^{-6}, 1 - 10^{-6}]$.

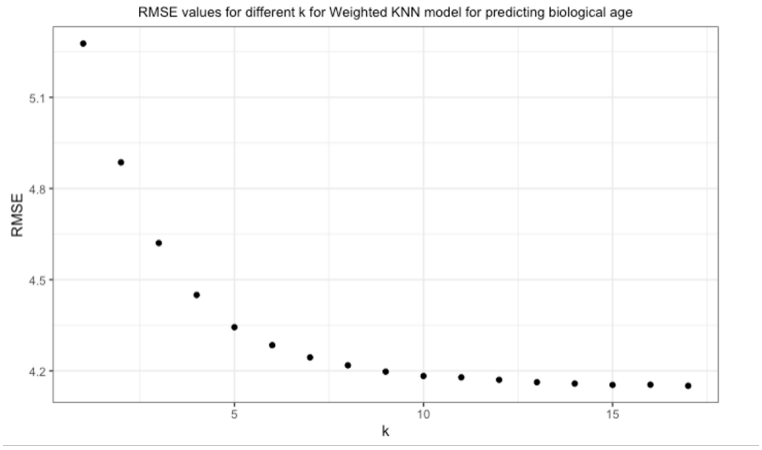

Figure 2: RMSE for different values of k

```
W <- D/maxdist
W <- pmin(W, 1 - (1e-06))
W <- pmax(W, 1e-06)
```

Now, the weights are further manipuated based on the kernel specified by the user. In our project, the "triangular" kernel gave us the optimal model. The following is the psuedo code for the manipulating the weights when the kernel is "triangular" :

```
if (kernel == "triangular")
        W <- 1 - W
```

Eventually, the weights are then fitted by first computing the sum of the products of each weight with each observed value in the neighbour and then dividing by the sum of weights:

```
fit <- rowSums(W * CL)/pmax(rowSums(W), 1e-06)
```

(Seinen, para. 1-3)

Based on our preliminary analysis, we also thought of using a model that is powerful enough to find hidden and complicated patterns. Hence, we decided to create a GBM model to predict the biological age.

We implemented H2O's GBM, a forward-learning ensemble method used for predictive analysis, by utilizing its following functions:

- Mean squared error as the loss function
- Stochastic gradient boosting by the use of column and row sampling

GBM within H2O obtains its predictive results in two steps -

1) Minimizing the model's loss function by utilizing the gradients.

2) Additively improves weaker models to increase the accuracy of predictions.

The Gradient Boosting Algorithm works in the following ways:

It begins by taking every row measurement and its corresponding response variable as input. It also takes in a differentiable loss function to evaluate our predictions. To predict biological age, we utilized Mean Squared error as our loss function.

Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable Loss Function $L(y_i, F(x))$

Then, we begin with our Step 1,

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

Here, $L(y_i, \gamma)$ defines the loss function. In our case, y_i represents the *Age at CT* and γ represents the predicted *Biological age*. In this case, we find the predicted value that minimizes the loss function, i.e. Mean Squared error. We can find such γ by simply taking the gradient. It is observed that the γ that minimized $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$ is the average of all the response values. Hence, $F_0(x)$ will just be a leaf that predicts each row as the same.

Now, we will loop M times to create M trees.

Step 2: for $m = 1$ to M :

In part A of Step 2, we just take the derivative of loss function with respect to the predicted value. Here, y_i is the Age at CT and $F(x_i$ is the predicted biological age. On simplifying this term, we observe that r_{im} is equal to residual (Observed - Predicted). We can calculate r_{im} for each sample by plugging in $F(x) = F_{m-1}(x)$.

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1, \ldots, n$

In part B of Step 2, we will be building a regression tree to predict the residuals. The terminal regions R_{jm}, for $j = 1 \ldots J_m$ are just the leaves of the regression tree.

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$

In part C of Step 2, we compute output for each j^{th} leaf in the tree such that it minimizes the summation mentioned below. In this case, the loss function also takes into account of the previous prediction ($F_{m-1}(x_i)$). Also, x_i defines the records that fall within a particular leaf. The output of each leaf is the average of the outputs of all records that fall in that leaf.

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

In part D of Step 2, we will make new prediction for each sample by utilizing the prediction of last step. In this step, we also use the learning rate to reduce the affect of each tree on the final prediction.

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum^{J_m} \gamma_{jm} I(x \in R_{jm})$

Finally, output the result in the end.

Step 3: Output $F_M(x)$

(Hastie et. al, 2001, p.339)

To optimize hyper-paramters such as number of trees, sample rate, etc. in the GBM, we used grids. To predict the biological age with only CT data, the GBM model was as followed:

```
h2o.gbm(2:12, 1, train, nfolds = 5, ntrees = 49,fold_assignment = "Modulo", score_tree_interval = 5,
max_depth = 3,min_rows = 1, sample_rate = 0.9, col_sample_rate = 0.7, max_abs_leafnode_pred = 100)
```

4.2.3 Pre-processing data to predict death using CT data

To decrease the variance of the reponse variable, i.e., 'Days Before Death', we converted this term into years by dividing the values by 365. Further more, in attempt to make the distribution non-uniform and less random, we took the log of the response variable.

4.2.4 Model description for predicting number of days before death

We again used Weighted K-NN and GBM to predict death using the CT scan data and newly computed biological age from the models mentioned above. To understand the underlying algorithms of Weighted K-NN and GBM please refer to section 4.2.2.

Using similar methods mentioned in section 4.2.2, the optimal value of "k" was 18, and the "kernel" used was "rectangular." When the "kernel" is "rectangular", it implies that no weights have been added and unweighted K-NN is used.

Likewise,to obtain optimal parameters we used Grids. The model parameters were:

```
h2o.gbm(c(1:2, 4:15), 3, train, nfolds = 5, ntrees = 24, max_depth = 3, sample_rate = 0.6,
col_sample_rate = 0.7,col_sample_rate_per_tree = 0.4)
```

4.3 Making predictions using CT and Clinical Data

4.3.1 Pre-processing data to predict biological age with clinical data

The pre-processing concept was similar to section 4.2.1. To filter out healthy people we only considered patients that had a BMI lesser than 30, consumed no tobacco, had a FRS 10-year risk score lesser than 0.19 (Bosomworth, 2011, para-6), and did not have any Metabollic synodrome. We removed outliers in FRAX scores and Aortic Calcification for each age group.

4.3.2 Predicting biological age with CT and clinical data

We used Weighted K-NN and GBM to predict the biological age using CT, clinical data. To understand the underlying algorithms of Weighted K-NN and GBM please refer to section 4.2.2. The Weighted K-NN model for predicting biolgical age had "k" equal to 7 and the "kernel" set to "triangular." Similarly, the optimal parameters for the GBM model was:

```
h2o.gbm(c(1, 3:16), 2, train, sample_rate = 0.8, col_sample_rate = 0.8,
ntrees = 44, max_depth = 10, nfolds = 5, max_abs_leafnode_pred = 90)
```

4.3.3 Predicting death with CT and clinical data

We used Weighted K-NN and GBM to predict death using CT, Clinical Data, and newly computed biological age in the section above. To understand the underlying algorithms of Weighted K-NN and GBM please refer to section 4.2.2. The parameters for Weighted K-NN were "k" equal to 23 and the optimial kernel found was "gaussian". The psuedo code for the "gaussian" kernel is:

```
if (kernel == "gaussian") {
        alpha = 1/(2 * (k + 1)) #Computes significane level
        qua = abs(qnorm(alpha)) #Finds quantile, or point in gaussian curve to represent alpha
        W = W * qua #Scale weights with "qua"
        W = dnorm(W, sd = 1) #Find probability density
    }
```

(Seinen, para. 1-3)

A significance value "alpha" is found based on the number of neighbours "k" used. Then, we find a point in the gaussian distribution representing "alpha" - here called as "qua". We scale the weights by then multplying them with "qua". Then the new weights are equal to the probability density of the scaled weights.

The optimal parameters for the GBM to predict death found using grids were:

```
h2o.gbm(c(1:2, 4:21), 3, train, sample_rate = 0.6, col_sample_rate = 0.7, ntrees =23,
max_depth = 8, min_split_improvement = 0.0001, score_tree_interval = 5, nfolds = 10)
```

5 Results

5.1 Predicting Biological Age

We initiated our experiment to predict *Biological age* using the Weighted KNN model on CT scan data. Using this model, we made predictions on how well it estimates the *biological age* of healthy people. With CT data only, the model's Root Mean Squared Error value came out to be **4.19 years**. After incorporating clinical data with CT data, we created the second model to predict *biological age* and the RMSE score dropped down to **2.6 years**.

Similarly, we predicted *biological age* using GBM. Using the CT data of healthy and alive people as training data, we created the GBM model which predicted the *biological age* with a RMSE of **3.28 years**. After incorporating clinical data with CT data, we created the second GBM model to predict *biological age* and observed that the RMSE score dropped down to **2.088 years**.

Figure 3 shows comparisons between the performance of Weighted K-NN and GBM models that were obtained after incorporating clinical data.

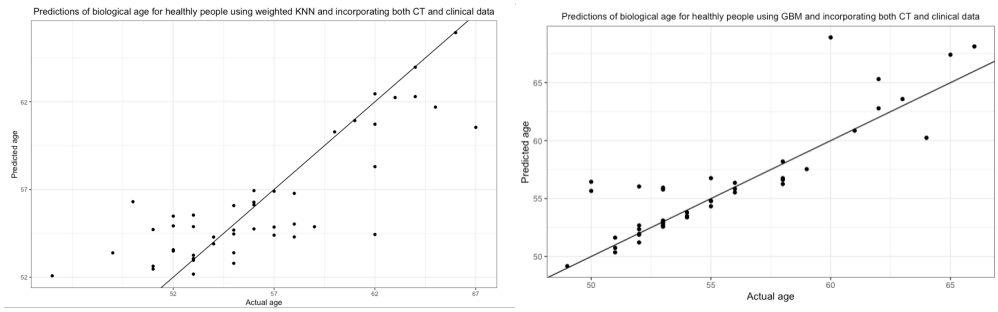

Figure 3: Comparing Biological Age predictions between Weighted K-NN and GBM

5.2 Predicting Death

When using only the CT data and predicted biological age, the Weighted K-NN model gives a RMSE of around **3.12 years**. On the other hand, the GBM model gives a RMSE of around **2.7 years**. Furthermore, the GBM recognizes that the Biological Age is the second most important feature to predict death; however, the GBM considers Muscle HU to be the most important feature to predict death.

Similarly, when incorporating clinical data, the weighted K-NN model gives a RMSE of around **2.4 years**. Likewise, the GBM models gives a RMSE of around **1.9 years**. Although Muscle HU and Biological Age seems to be the most important features to predict death, but FRAX scores have become the third most important feature.

From the UMAP visualizations in Figure 4 we can see how our best model, i.e, GBM with clinical data makes predictions that are similar in nature to the actual data

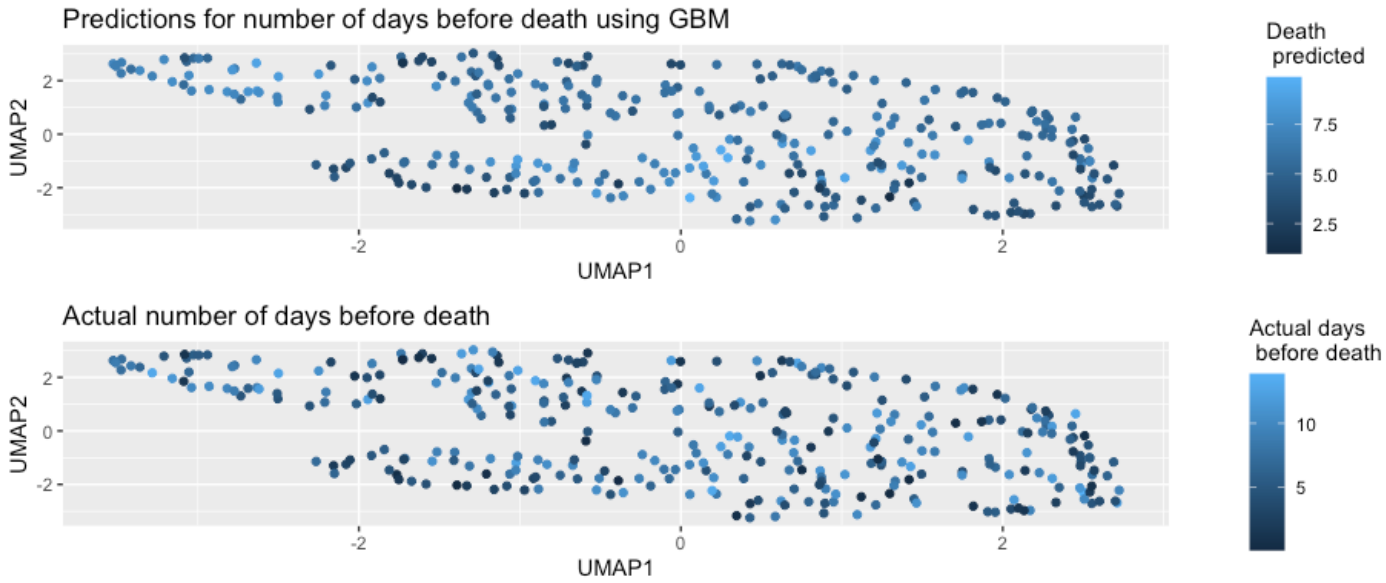

Figure 4: Comparing death predictions made by the best model with actual dataset

6 Future Work

To improve the Accuracy model, we could probably use a Stacked Ensemble with multiple GBMs, Random Forests, and Deep Learning models. We could perform Hypothesis testing on varibles such as Tobacco against death to verify or understand why it was not the most important feature. We could receive more data about healthy patients to improve the biological age predictions.

7 References

Seinen, D. (n.d.). Weighted K-nearest neighbors and R kknn package. Stack Overflow. Retrieved May 12, 2022, from https://stackoverflow.com/questions/65654487/weighted-k-nearest-neighbors-and-r-kknn-package

Bosomworth, N. J. (2011, April). Practical use of the Framingham Risk Score in Primary Prevention: Canadian Perspective. Canadian family physician Medecin de famille canadien. Retrieved May 12, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3076470/

Hastie, Trevor, Robert Tibshirani, and J Jerome H Friedman. . (2001). In The Elements of Statistical Learning. (Vol. 1, p. 399). essay, Springer New York.

Mayo Foundation for Medical Education and Research. (2022, January 6). CT Scan. Mayo Clinic. Retrieved May 12, 2022, from https://www.mayoclinic.org/tests-procedures/ct-scan/about/pac-20393675

Pickhardt PJ;Graffy PM;Zea R;Lee SJ;Liu J;Sandfort V;Summers RM; (n.d.). Automated CT biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: A retrospective cohort study. The Lancet. Digital health. Retrieved May 12, 2022, from https://pubmed.ncbi.nlm.nih.gov/32864598/