CS 760: Machine Learning Spring 2022

CS760 MACHINE LEARNING PROJECT

AUTHORS: Pankaj Kumar, Lindsay Noelle Heimerl, Krishna Chaitanya Goparaju

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

1 Abstract

This project focuses on the concept of the biological age of an individual and building an efficient model for its prediction. This is achieved by using modern Machine Learning algorithms on a dataset provided by the Department of Radiology, University of Wisconsin-Madison. This dataset is unique in the sense that it not only contains regular medical test results like Computer Topography Data but also has columns related to adverse clinical outcomes like cancer, death, Alzheimer, diabetes, etc which helped in accurately modeling biological age. For the prediction of biological age, we have used classical Machine Learning models like Linear Regression, Decision Trees, and Random Forest (Goal 3). We trained our model on the given dataset and did a comparative analysis of the above ML models. We also studied the impact of augmenting our CT Data with demographic info(sex, age, etc) and unhealthy habits (tobacco, alcohol) for predicting biological age. Lastly, we extended the biological age concept to find adverse clinical outcomes like time to death, etc using the same ML models(Goal 1 and Goal 2). We have used open-source python packages like Scikit-Learn, Pandas, etc for our implementation purposes.

2 Introduction

Human aging is a gradual and complex process that is dependent on different types of tissues. Aging leads to diseases and the malfunction of various organs over time. The majority of the biological markers that are used to characterize age differ greatly from person to person even with the same chronological age. However, biological age is a conceptual idea that can be used to tackle aging-related issues and is believed to be the true age of the individual. Thus, biological age provides a better measure for the life expectancy of an individual than his or her chronological age.

The main idea of this project is to make use of unused data from regular medical tests alongside an individual's day-to-day habits and feed them to a prediction model to predict the biological age of a person. In addition, we extended the knowledge of biological age to predict adverse clinical outcomes of individuals.

3 Related Work

A great deal of work has been done on our research topic of predicting the biological Age of any individual. Nowadays, we encounter many researchers who are using Machine Learning techniques for this purpose. A good application of this can be seen in the work of Xingqi Cao [1], and his colleagues where she used blood biomarkers as input features

to their ML models for the estimation of Biological Age. They used the R-squared value for their model evaluation and found a gradient-boosting regressor algorithm as their best model.

Another interesting work is done by Wang[2] and his colleagues who developed a composite biological age predictor having a high correlation with chronological age. They used an extreme gradient boosting (XGBoost) algorithm and trained it on a multisystem measurement dataset collected from the Dongfeng-Tongji cohort study in China. They also did a feature importance analysis to find the most important health factors in predicting biological age. We have also seen the usage of CoX models in [3], a well-known statistical method to associate the biological age of an individual to a DNA methylation-based predictor measures/patterns.

Therefore, the applications of Machine Learning and data-analysis applications in the above studies give us the confidence of applying Machine Learning models to our dataset and getting a good model for biological age.

4 Dataset

The dataset used was provided by Prof. Perry Pickhardt from the Department of Radiology at UW-Madison. The dataset contains a total of 52 features which is a mixture of patient's demographic, medical history, and measurements from the medical test which generally gets overlooked, but could prove to be of use during medical model development.

Now, talking more about the dataset, it can be subdivided into the following three sections. The first section also referred to as "Clinical Data" specifically consists of patient's demographics like sex, age, patient BMI, smoking/drinking habits, and risk scores (FRS) for cardiovascular and hip-related diseases. The second section, labeled "Clinical Outcomes", has data related to the occurrence of negative health diagnoses such as diabetes, Alzheimer's, cancer, etc. The dataset also contains a prominent feature "Death" which refers to the number of days a patient died after the initial computerized tomography scan, which we have used for developing our adverse outcomes model in future sections.

Lastly, we have the "Computer Tomography" data. In short, computed tomography, also well-known as CT Scan data is collected by using both standard X-Ray examinations and computer technology to generate images of the body parts. In the dataset used, the input features in the CT data section contain results from body parts like bones, muscles, fat, and heart. The below figure contains a sample example of what CT-SCAN images look like.

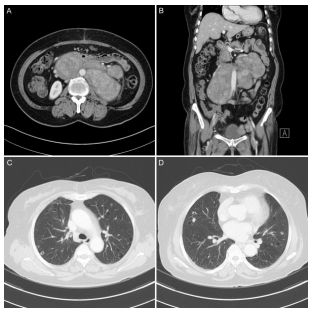

Figure 1: CT-SCAN

Apart from that,this dataset is a real world dataset, therefore it comes with the problem of missing values. This could have happened due to many reasons like data entry error, human errors etc. A quick statistical analysis across sections shows that CT Data has in total 964 missing values which is dwarfed by the 11258 missing values count in Clinical Data section. The next section shows the data handling done for these features of our dataset.

5 Approach

Our approach to research was broken down into two separate tasks: initial data cleaning and pre-processing data for the main research goals.

5.1 Data Cleaning

For the initial data cleaning, each section of data was individually processed based on types of features. These separate partitions of data were defined as the "Clinical Data", "Clinical Outcomes", and "Computerized Tomography Data" information sections described previously. For each of these sections, features were analyzed for the distribution of their values if the feature showed numerical data types and the number of missing values.

The "Clinical Data" had a mix of classification features (such as those documenting sex, alcohol abuse, tobacco use, etc) that were changed from string-based data types to a boolean scheme. The remaining features were numerically based, with some having a subset of missing values ("nan" values). To ensure no bias was introduced into the dataset, features with below a fifteen percent threshold for quantity of "nan" values had those missing values replaced with the average for the column. Features with a higher value of missing inputs were not used. Lastly, all non-boolean features were scaled using the StandardScaler package, which transforms numerical features by removing the mean and scaling the input to unit variance. Below is a mathematical representation of this transformation to the data. Here, x represents the input data, with mu and s representing mean and variance respectively.

$$z = (x - \mu)/s$$

This scaling is done to ensure the performance of models, as they may underperform if data is not aligned with a standard normal distribution. As we are using mean squared error as the main metric for comparing the performance of our models, having our data scaled to the same range enables a more accurate evaluation.

The "Clinical Outcomes" data features focused on a few specific negative health diagnosis (cancer, diabetes, Alzheimer's, etc..) and the date of the diagnosis after the initial computerized tomography data (ct) was taken. While the majority of feature values were missing, data was only entered in if the sample/patient had a recorded diagnosis, which lead to missing values being understood to be the lack of a diagnosis. This meant that each feature indicating a set of diagnosis was transformed into a boolean, while features describing when the diagnosis was given after the initial computerized tomography scan was dropped.

Lastly, the "Computerized Tomography" data section was focused on measurements of the body. While some features did present missing data, the data gaps were sparse, leading to them being replaced with the average of the column. The features were lastly scaled using the same method as applied with the clinical data to ensure that if the ct data was used with the rest of the clinical data the values would overall have the same standard normal distribution.

5.2 Pre-Processing

The approach to pre-processing the data was to be able to make the data into simple classification or regression problems for each individual goal.

For adverse health outcome prediction, we used a feature from the clinical outcomes data section which had information on how many days the individual took to pass away after the initial ct scan. Missing data within the feature indicated that the patient had not passed. We divided days given for passed patients by the number of days in the year, changing the focus from predicting days until death to predicting years until death post ct scan. For individuals who had not passed we estimated the year they were likely to pass (80 years based on human averages) and used that value

instead. This altered death feature was used as the label for our classification/regression models in both our goals 1 and 2.

Our pre-processing for mapping biological age started with applying our clinical outcome data. Biological age takes into account the health state of the individual as well as their chronological age, which leads to us combining both of those factors to develop a rough estimate for each individual. The number of adverse health outcomes for each individual was added together and scaled by a weight based on the quantity. The scaling applied to the adverse outcomes progressively increased with higher amounts of diagnosis, with zero health outcomes having scaled by zero and increasing to 7 scaled by 4. After, the chronological age of the individuals was added to this scaled sum of health issues. We can see the proposed Biological Age vs Chronological Age in Figure 2.

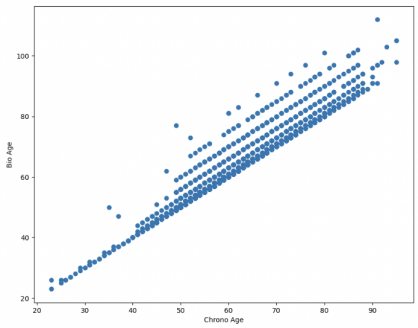

Figure 2: BiologicalAge vs ChronologicalAge

From the above plot, we can infer that this formulation is synchronous to the real-life fact that older individuals with more number of adverse health issues having higher biological age as well as younger individuals with less or no adverse outcomes just having a biological age representation of their chronological age. This combination of health data was used as the label for classification models within goal 3.

5.3 Algorithms

In our project, the first problem that we tackled is developing a model which accurately predicts the biological age of a patient. For our experiments, we are using the following models

- Linear Regression: It is no doubt that Linear Regression is perhaps the most well-known and well-studied algorithm in the field of Machine Learning. In simple terms, Linear Regression assumes a linear relationship between a set of inputs and their output. Following is an example of a simple linear regression problem. Now, for our use case, We are trying to solve a multiple linear regression problem since we have more than one input variable (CT Data and Clinical data) and a single output(bio-age).

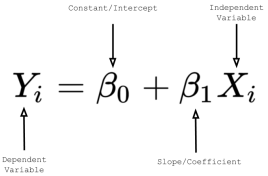

Figure 3: Linear Regression [4]

- Decision Trees: As the name suggests, the Decision trees form a regression model which resembles the structure of a tree. The tree consists of nodes that are of two different types. First is an internal node which is related to a feature and branches out to two or more nodes. The second node type is called the leaf node which contains the predicted value for the output. For any new data, the prediction is given by following the path in the decision tree model, starting from the first internal node (also called the root node) to the leaf node which gives the prediction for the new data.

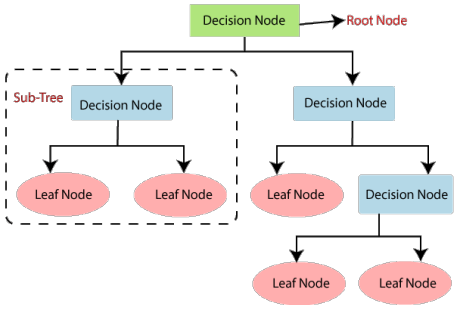

Figure 4: Decision Tree [5]

- Random Forest: This algorithm comes under the family of ensemble methods used in Machine Learning. Unlike individual models discussed above, an ensemble runs multiple machine learning models and combines their predictions. In our case, Random Forest consists of several decision trees for training, and then it outputs the mean of all the predictions from each decision tree.

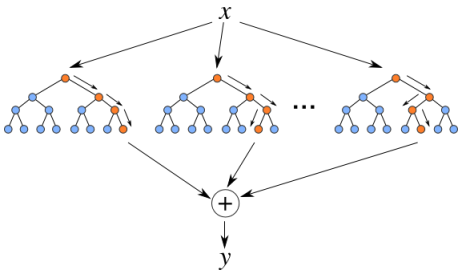

Figure 5: Random Forest [6]

5.4 Packages

Figure 6: Key Packages used []

For the purpose of implementing the above algorithms, we have used ”Scikit-Learn”, which is a well-known library that provides efficient tools for predictive data analysis. It provides various supervised learning algorithms for classification/regression problems and clustering algorithms (KNNs, SVM, etc) for unsupervised learning. We have also

heavily used the ”Pandas” library for data cleaning and pre-processing steps. Apart from that, libraries like NumPy for mathematical manipulation, matplotlib, plotly, and seaborn libraries for generating plots for visualization purposes were also used.

6 Results

6.1 Experimental Setup

For our experiments, as given in the deliverables, We have divided them into two parts. In the first part, we are training our model with just the Computerized Tomography Data i.e (CT Data). For the second part, we are augmenting the CT Data with the Clinical data too. For all our experiments, we are using 70 percent of the data for training and the rest 30 percent for testing purposes. Since we are predicting an individual's bio-age which is numeric data, we are using Root Mean Squared Error (RMSE) as our metric for model comparison. Following is the definition of the RMSE error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Figure 7: RMSE : Root Mean Squared Error

6.2 Biological Age Prediction

For goal 3, as given in the project description, we are predicting the biological age of any individual. The bar graphs in figure 7 contain the results for both above-mentioned experiments for our biological age prediction models. Please note that the blue bar is for CT Data and the Green bar is for both CT Data and Clinical Data.

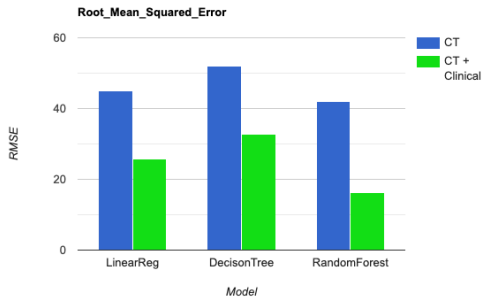

Figure 8: Biological-Age Prediction

6.3 Adverse Clinical Outcome

For achieving goals 1 and 2, we are predicting adverse clinical outcomes. For that, as mentioned in the approach, we are trying to predict the approximate time frame for their death. The bar graph in Figure 8 shows the result.

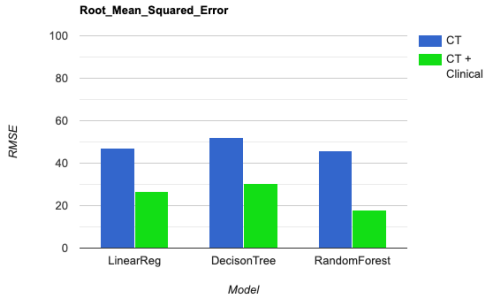

Figure 9: Clinical Adverse Outcome: Death

6.4 Feature Importance

After using the above Machine Learning models, we also tried to plot the feature importance of our dataset containing both CT Data and Clinical Data. The pie chart given in Figure 9 shows the feature importance we got:

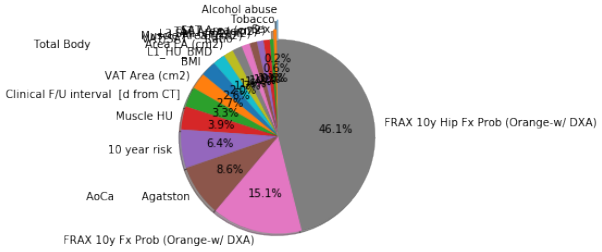

Figure 10: Feature Importance

For the calculation of feature importance, since we achieve the best performance on the Random Forest model, we are directly associating the weights given to each feature in the model with its importance value.

6.5 Conclusion

Across all the goals and the different plots that we have obtained for all the three models, we can deduce the following conclusions:

1. The addition of the clinical data as well as the ct data being used for training/testing lead to an increase in performance shown across all goals (seen as a decrease in the mean squared error in the results section).
2. Among all the models, Random Forest performed the best, followed by Linear Regression and Decision Tree in order. The supremacy of random forest over other models used could have been attributed to the average of multiple decision trees being used. Select features from the dataset have more influence over the classification outcome, so a number of trees splitting based on different features results in better outcomes than a single tree. Multiple trees within a random forest likely are splitting on more influential features early on, resulting in a lower mean squared error and higher performance.
3. Regarding features, we can infer from the pie chart that risk scores(FRAX) from clinical data turned out to be most important once followed by ”AoCa Agatston” from CT Data. This also proves the importance of using

both sections of the dataset to get better models for predicting Biological Age.

6.6 Future Work

For potential future works, more research could be done on the impacts of the adverse health outcomes covered within the clinical outcome data. Likely, not all diagnosis shown there would have the same impact on the biological age of the individuals impacted. While our model for goal 3 has them impacting biological age to the same degree, each health outcome should be scaled to their actual impact on lifespan. Likewise, time of death prediction for goal 1 could be improved by incorporating information from adverse clinical health outcomes into the death year label. Individuals with no death recorded should have lower death time labels if they have a recorded diagnosis than those without a death recorded and without any recorded adverse health outcomes.

7 Github Repostiory

Please find the (Github Repository) which contains the code for our models.

References

[1] Xingqi Cao, Guanglai Yang, Xurui Jin, Liu He, Xueqin Li, Zhoutao Zheng, Zuyun Liu, and Chenkai Wu. A machine learning-based aging measure among middle-aged and older chinese adults: The china health and retirement longitudinal study. *Frontiers in Medicine*, 8, 2021.

[2] Chenming Wang, Xin Guan, Yansen Bai, Yue Feng, Wei Wei, Hang Li, Guyanan Li, Hua Meng, Mengying Li, Jiali Jie, Ming Fu, Xiulong Wu, Meian He, Xiaomin Zhang, Handong Yang, Yanjun Lu, and Huan Guo. A machine learning-based biological aging prediction and its associations with healthy lifestyles: the Dongfeng-Tongji cohort. *Ann. N. Y. Acad. Sci.*, 1507(1):108–120, January 2022.

[3] Brian H Chen, Riccardo E Marioni, Elena Colicino, Marjolein J Peters, Cavin K Ward-Caviness, and Tsai.. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)*, 8(9):1844–1865, September 2016.

[4] Aiden Wilson.LinearRegression: Towards Data Science. `shorturl.at/muMS5`.

[5] DecisionTree:Javatpoint . `shorturl.at/ptW48`.

[6] Jagandeep Singh(Medium). RandomForest: Pros and Cons . `shorturl.at/eBMQ6`.