CS 760: Machine Learning Spring 2022

Estimating adverse clinical outcomes and Biological Age using CT and Clinical Data

AUTHORS: Ganesh Cheerla(gcheerla), Harsh Sahu(hsahu), Hemalkumar Ratilal Patel(hrpatel5)

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

We have a real dataset related to *Opportunistic Cardiometabolic* Screening. The goal is to exploit the dataset collected for the mentioned reason to predict other clinical outcomes and measure biological age using incidental data that typically goes unused/underuse – this includes clinical data and Computerized Tomography (CT) data. For the context of this project, we have predicted death, diabetes and heart-attack and derived biological age.

1 Introduction

Our goals for this project are as below:

1. Predict Clinical Outcomes using first only CT data, and then CT+Clinical data. Later we compare both outcomes to assess the effectiveness
   (a) Death
   (b) Diabetes
   (c) Heart-Attack
2. Derive a patient's Biological Age

For (1), we have split the data into train/test and considered relevant key clinical outcome column for that particular prediction, i.e., *DEATH [d from CT]* for Death, *Type 2 Diabetes DX Date [d from CT]* for Diabetes and *MI DX Date [d from CT]* for Heart Attack.We have used 2 approaches for prediction, which we will describe in **Section 4**.

For (2), we have only considered CT data as an input features. CT data contains columns that are proven biological age markers and have been used widely in clinical research [1] [2] [3] already. We have formed train data using patient who have died, and rest are grouped into test data to calculate biological age. This is because we know for certainty that patient who died reached their biological limit. The exact methodology is explained in **Section 4**

2 Related Work

There are studies that analyze a CT feature and examine its importance in predicting adversarial outcomes like death or critical illness. [5]analyzes the impact of Bone Mineral Density of L1 vertebra. It mentions the

values ranges of < 100 to be osteopric which we try to validate in our experiments. Similarly, [4] analyzes the impact of various fat measures like Visceral Adipose tissue area and Subcutaneous Adipose tissue area on health and metabolism of patients. We use these observations to get patterns from our data using different ratios like Total adipose tissue area/Body area, VAT/SAT. Aortic calcification is a measure of calcium deposit in heart's blood vessels. As per common medical knowledge, it is a best indicator of Heart Attack. We analyze the work about this feature [7], and confirm through our experiments. Liver HU indicates the attenuation of liver fat. The lower this value, the higher the risk of fatal outcomes like death. This is anlyzed in [6]

There have been clinical researches [1] [2] [3] to exploit biological health markers to effectively calculate biological age. All these studies have incorporated different techniques such as PCA, Linear Regression, multiple linear regression (MLR), Klemera and Doubal's method (KDM) etc. All these studies collected data on healthy individuals periodically to assess how biological health markers get affected as the chronological age grows. They have considered them a baseline and then use that to calculate biological age of test samples.

However, we don't have the same type and amount of data as those studies. So we cannot use them as it is. However, the inspiration is that biological health markers are effective at measuring biological health.

3 Dataset

The dataset was made available by Perry Pickhardt (Department of Radiology, UW-Madison) who also provided guidance in supervising and evaluating this project. The dataset contains 3 set of columns.

- Clinical Data
  - Cols A-C: anonymized Case ID info
  - Col D: Clinical F/U interval [days from CT]
  - Cols E-J: pt BMI, sex, age (at time 0=CT date), smoking/drinking hx)
  - Col K: FRS = Framingham Risk Score (multi variable 10-yr cardiovascular risk score)
  - Cols L-M: FRAX = Fracture risk assessment score (multi-variable 10-yr risk for all and hip fx)
  - Col N: Metabolic Syndrome (Y/N/blank=unknown) ? really more of an outcome
- Clinical Outcomes
  - Col P: Death
  - Cols Q-V: Cardiovascular events w/ dates (CVD=stroke, Heart failure, MI=heart attack; any=positive)
  - Col W-X: T2 Diabetes (if dx)
  - Cols Y-AH: Pathologic/osteoporotic fracture w/ date (any=positive; femoral=hip fx)
  - Cols AI-AJ: Alzheimer's Dx
  - Cols AK-AN: Cancer Dx's
  - Col N: Metabolic Syndrome; could be considered an outcome)
- Computerized Tomography Data
  - Col AP: Bone measure/BMD (L1 HU)
  - Cols AQ-AU: Fat measures (total/visceral/subcutaneous; V/S ratio; all total body X-section)
  - Cols AV-AX: Muscle measures (HU/Area/SMI)
  - Col AY: Aortic Calcification (Ag)
  - Col AZ: Liver fat (HU)

4 Approach

4.1 Packages

We have extensively used python machine learning and data-science libraries throughout this project.

- *numpy* and *pandas* are used to handle dataset and perform operations
- *seaborn* and *matplotlib* are used to visualize the dataset and output results
- *sklearn* and *xgboost* are used to run machine learning algorithms

4.2 Data Cleaning and Pre-processing

We have used data clipping to remove skewness in features. Columns *AoCa Agatston* and *Liver HU (Median)* are clipped at 99th percentile.

Moreover, the sparse nature of the dataset prompted us to fill *NULL* values in CT data using the corresponding mean.

We have also used the iterative imputing strategy [8] to fill null values in a few CT features. This type of imputing is useful when we know that a feature is dependent on other features. In this dataset, we have filled the NULL values in CT features using dependent clinical features. First, we get the correlation heatmap of all the features. Using this, we determine the features on which a CT feature is dependent. We will use these dependent features as related columns and apply *sklearn.IterativeImpute*r. Under the hood, iterative imputer fits a linear regression model on non NULL values of related clinical features as 'features' and the CT feature as the 'target'. This ensures a more realistic NULL values as opposed to filling everything just by mean. For example, we fill the NULL values in 'VAT/SAT ratio' using 'Muscle_Area' and 'Sex' features. We have also used correlation heatmap 1 to drop features that are highly correlated and to add new features. We have removed TAT area and Total_body_area and added a new feature 'TAT area/Total_body_area'

4.3 Methodology - Predicting Death, Diabetes and Heart Attack

For each 3 predictions, we first use only CT data, and then take the results again with CT + Clinical data. Our experiment involves running 3 algorithms, and comparing their corresponding RMSE(Root Mean Square Error)

1. Linear Regression
2. Support Vector Regressor
3. XGBoost

For each 3 predictions, we split the original dataset of size 9223 items in 80-20 train/test fashion. It includes both *NULL* and *NON-NULL* key columns. For example, for column *DEATH [d from CT]* we have

- Train Data - 7363, 6944 NULL, 419 NON-NULL
- Test Data - 1860, 1730 NULL, 130 NON-NULL

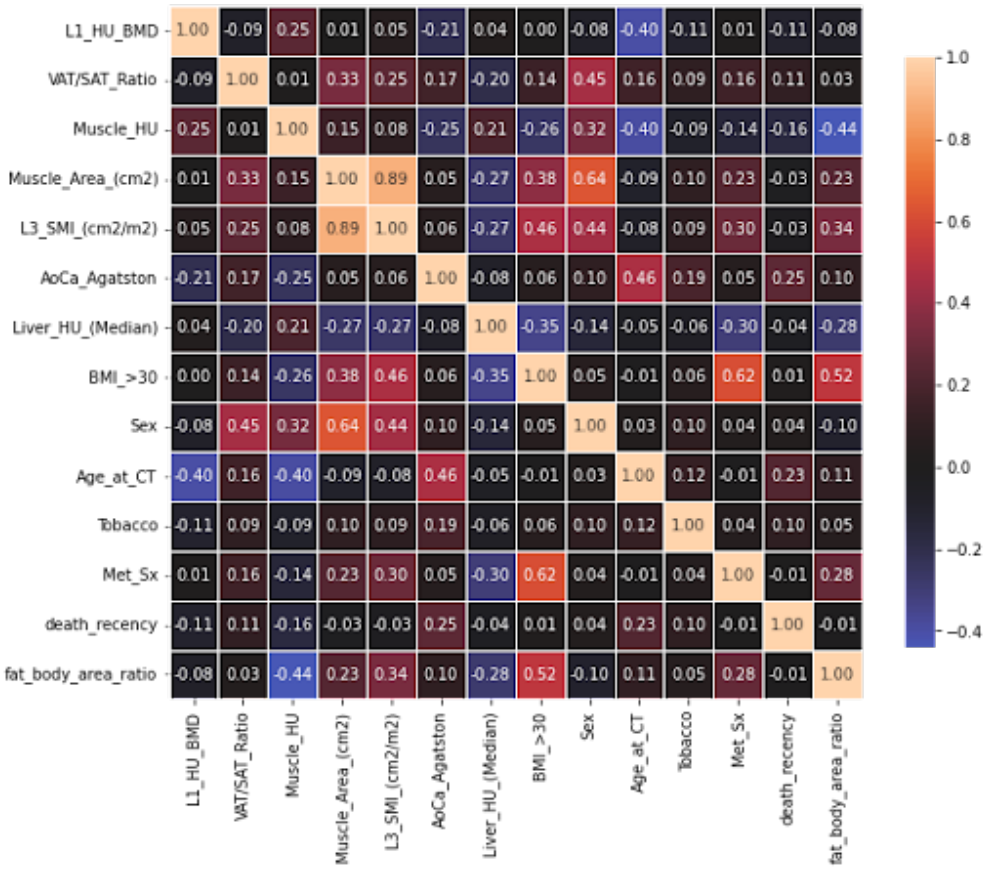

Figure 1: Feature Heatmap

4.3.1 Approach-1:

We don't perform any other modification or data pre-processing in this approach.

For each 3 predictions, we perform the following operations:

1. We train our model **only on NON-NULL training data** and test on NULL test data.
2. We compare the distribution of values computed in the previous step with that of the NON-NULL test data

4.3.2 Approach-2: Recency

In this approach, we introduced a new column *recency*.

$$death\_recency = 1/DEATH[d\_from\_CT]$$

$$diabetes\_recency = 1/DX\_Date[d\_from\_CT]$$

$$heart\_attack\_recency = 1/MI\_DX\_Date[d\_from\_CT]$$

This gives us the chance to fill all NULL values with zeros. However, it results in skewed data - Fig 2. Hence, we transformed non-zero values (using Box-cox transform) to a more uniform distribution. Later, we sub-sampled zero samples to be equal to non-zero samples. This provides us with different training/test number of samples than the original. For example, for *death_recency* we got 838 training samples and 1860 test samples.

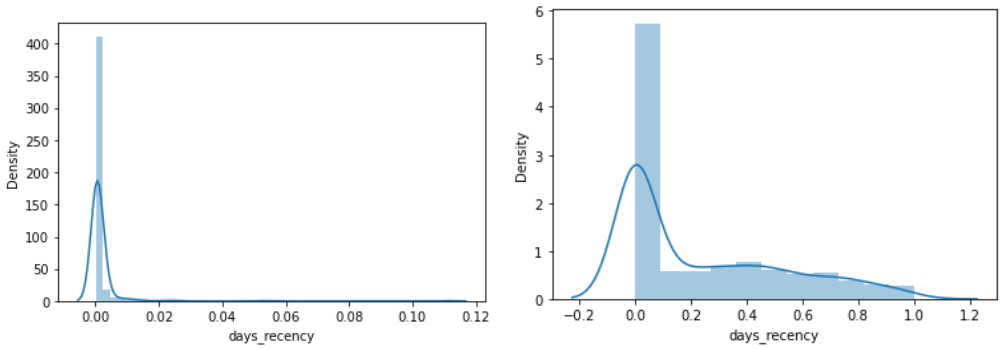

Figure 2: left - skewness in recency, right - skewness made better

For each 3 predictions, we perform the following operations:

1. We train our model on the **entire training data** and test on NULL test data.
2. We compare the distribution of values computed in the previous step with that of the NON-NULL test data

4.4 Methodology - Deriving Biological Age

We split the dataset into train and test using *DEATH [d from CT]* column. The data item with non-empty value will go into training data, and rest will be used to check the effectiveness of the methodology. Based on this, we split the original dataset of size 9223 into 549 training items and 8674 test items.

We made few intuitive assumptions:

1. People die when they hit biological age 100. Healthy people can slowdown the bio age and that is how some humans live more than 100 chronological age. This gives us $max\_bio\_age\_in\_days = 36500$
2. Higher the *DEATH[d from CT]*, higher the patient has *bio_days_left* to live

Based on this assumption, we define the biological age for training data as the function of *bio_days_left*.

$$bio\_age = max\_bio\_age\_in\_days - bio\_days\_left$$

$$bio\_age\_left = A + C * (1 - \exp^{-k*x})$$

Here, *bio_days_left* follows exponential decay function(increasing) form as shown in Fig - 3

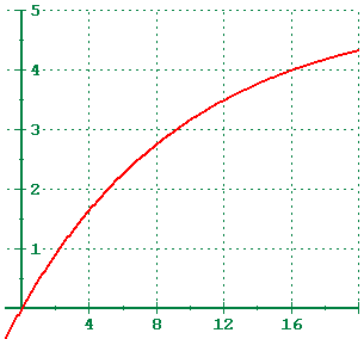

Figure 3: Example of a exponential decay increasing form plot

We calculate *bio_days_left* for training data, and then apply 2 the following 2 algorithms

1. Linear Regression
2. XGBoost

5 Results

5.1 Predicting Clinical Outcomes

For both the approaches mentioned in **Section 4** we executed 3 machine learning algorithm mentioned in **Secion 3**.

For approach-1 and *death*, we got the RMSE comparison results using only CT and CT + Clinical data as described in Table 1. XGBoost performed better than its counterparts with both CT and CT + Clinical Data and exhibited the lowest RMSE.

Model	**RMSE (only CT)**	**RMSE (CT + Clinical)**
Linear Regression	1351.75	1324.57
Support Vector Regressor	1410.32	1381.12
XGBoost	**1331.59**	**1314.85**

Table 1: Predicting Death Approach-1:Model Comparison

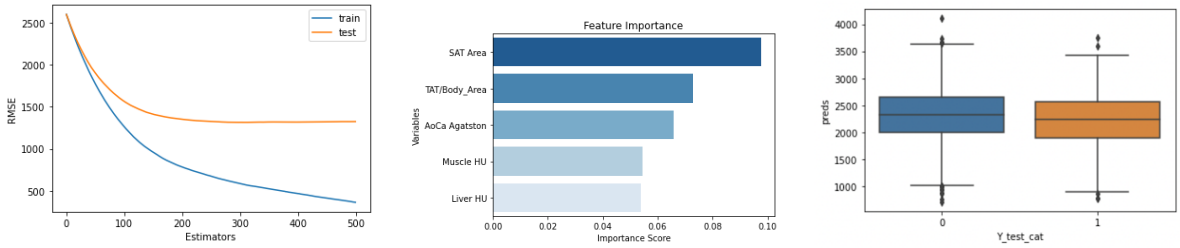

Figure 4: XGBoost - Predicting death without recency(approach-1). left = decreasing RMSE with number of iterations, center = feature importance, right = distribution comparison with test-data

For approach-2 and *death*, we got the RMSE comparison results using only CT and CT + Clinical data as described in Table 2. Again, XGBoost performed better than its counterparts with both CT and CT + Clinical Data and exhibited the lowest RMSE.

Model	**RMSE (only CT)**	**RMSE (CT + Clinical)**
Linear Regression	0.235	0.204 (+13.1%)
Support Vector Regressor	0.235	0.243 (-3.4%)
XGBoost	**0.232**	**0.212 (+8.6%)**

Table 2: Predicting Death Approach-2:Model Comparison

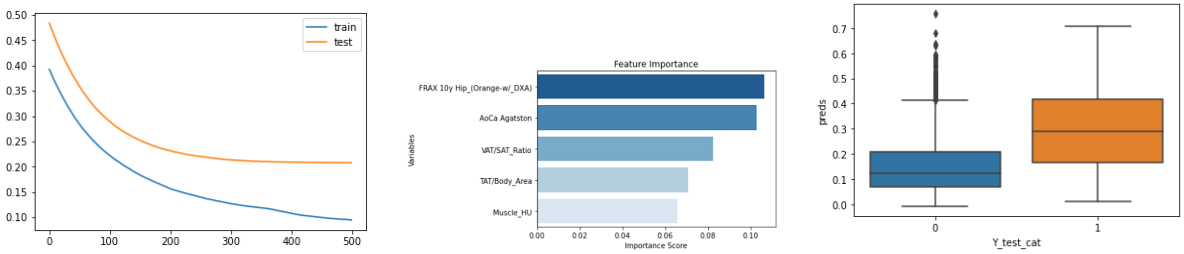

Figure 5: XGBoost - Predicting death with recency(approach-2). left = decreasing RMSE with number of iterations, center = feature importance, right = distribution comparison with test-data

We compared the results from both the approaches with the distribution of death, and found out that distribution in approach-1 result is almost same as test-data. However, with approach-2, it is slightly higher(lower because recency is inverse of days). Hence, approach-2 is better. The approach-1 output SAT Area, TAT/Body Area, AoCa Agatston, Muscle HU, Liver HU as the important features. On the other hand, the approach-2 considers FRAX, AoCa Agatston, VAT/SAT Ratio, TAT/Body Area, Muscle HU. Considering both the approaches, it seems that *AoCa Agatston, TAT/Body Area, Muscle HU* are important features.

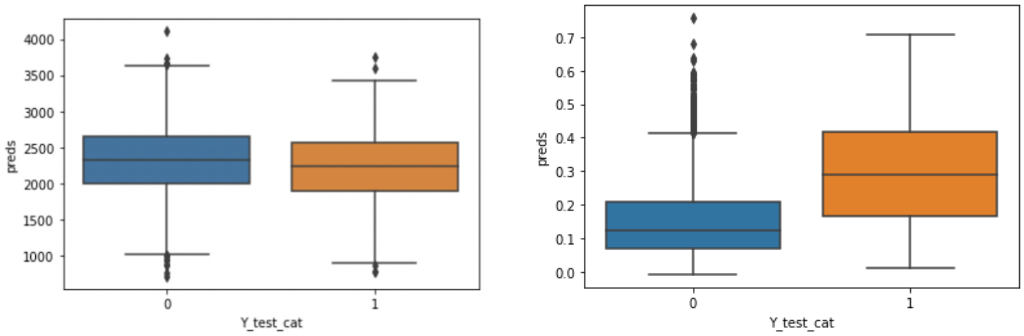

Figure 6: Comparison of both the approaches with test data on death distribution. left=approach-1, right=approach-2

We continued the same experiments with other clinical outcomes, and found out that *recency* approach is effective in the given dataset.

For approach-2 and *diabetes*, we got the RMSE comparison results using only CT and CT + Clinical data

as described in Table 3. Linear Regression performed better than its counterparts with both CT and CT + Clinical Data and exhibited the lowest RMSE.

Model	**RMSE (only CT)**	**RMSE (CT + Clinical)**
Linear Regression	**0.248**	**0.243 (-2%)**
Support Vector Regressor	0.253	0.258 (+1.9%)
XGBoost	0.257	0.251 (-2.3%)

Table 3: Predicting Diabetes Approach-2:Model Comparison

For approach-2 and *heart attack*, we got the RMSE comparison results using only CT and CT + Clinical data as described in Table 3. Linear Regression and XGBoost both performed better than Support Vector Regressor with both CT and CT + Clinical Data and exhibited the great improvements.

Model	**RMSE (only CT)**	**RMSE (CT + Clinical)**
Linear Regression	0.223	**0.221 (-0.8%)**
Support Vector Regressor	0.243	0.241 (+0.8%)
XGBoost	0.245	**0.230 (-6.1%)**

Table 4: Predicting Heart Attack Approach-2:Model Comparison

5.2 Deriving a Biological Age

We executed the approach mentioned in the **Section - 4** on both *Linear Regression* and *XGBoost* algorithms. As shown in the Fig - 7 Linear Regression performs worse and is concentrated around somewhere between 75-85 for most test samples. However, XGBoost shows Gaussian distribution and scattered nicely, which shows better fit.

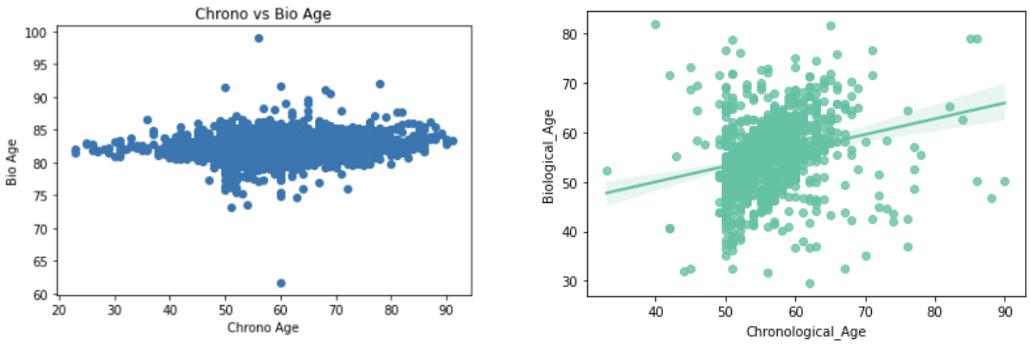

Figure 7: Biological Age: comparison of both the algorithm on test-data left=Linear Regression, right=XGBoost

Next, we show in Fig - 8 correlation of features such as muscle mass, muscle area and calcification with Biological age. Muscle Mass and Muscle Area are decreasing as Biological age increases. Calcification is increasing as Biological age increases.

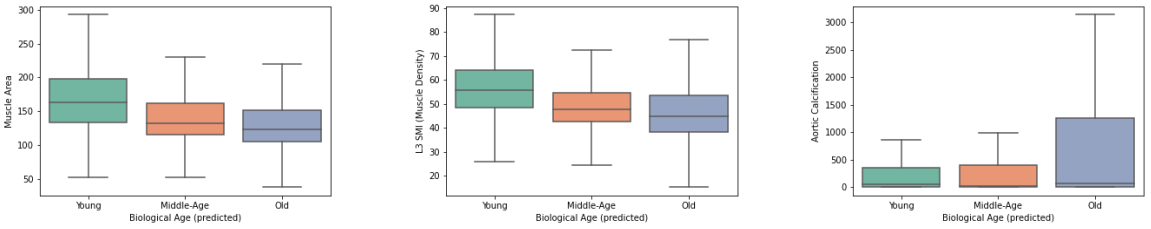

Figure 8: Effect on body as bio age is increasing. left = decreasing muscle area, center = decreasing muscle mass, right = increasing calcification

6 Conclusion

We shown in the project that it CT + Clinical data is better at predicting critical clinical outcomes such as *death, diabetes and heart-attack*. Moreover, CT data contains effective biological health-markers that can be used to assess how is patient doing biological health-wise. Although limited, for this dataset, XGBoost seems to be performing well in most cases.

One can find the entire project on GitHub [9]

7 Aknowledgement

We would like to thank Prof Danile Pimental (Department of Bio Statistics, UW-Madison) and Perry Pickhardt (Department of Radiology, UW-Madison) for their constant support and guidance in our endeavour. It would not be possible without their encouragement and constant quest of improvisation.

References

[1] Jia L, Zhang W, Chen X. Common methods of biological age estimation. Clin Interv Aging. 2017;12:759-772. Published 2017 May 11. doi:10.2147/CIA.S134921

[2] Levine ME, Crimmins EM. Is 60 the New 50? Examining Changes in Biological Age Over the Past Two Decades. Demography. 2018;55(2):387-402. doi:10.1007/s13524-017-0644-5

[3] Jia L, Zhang W, Jia R, Zhang H, Chen X. Construction Formula of Biological Age Using the Principal Component Analysis. Biomed Res Int. 2016;2016:4697017. doi:10.1155/2016/4697017

[4] Shuster A, Patlas M, Pinthus JH, Mourtzakis M. The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. Br J Radiol. 2012;85(1009):1-10. doi:10.1259/bjr/38447238

[5] Yaprak, G., Gemici, C., Seseogullari, O. O., Karabag, I. S.; Cini, N. (2020). CT derived Hounsfield Unit: An easy way to determine osteoporosis and radiation related fracture risk in irradiated patients. Frontiers in Oncology, 10. https://doi.org/10.3389/fonc.2020.00742

[6] Hamer, O. W., Aguirre, D. A., Casola, G., Lavine, J. E., Woenckhaus, M.; Sirlin, C. B. (2006). Fatty liver: Imaging patterns and pitfalls. RadioGraphics, 26(6), 1637–1653. https://doi.org/10.1148/rg.266065004

[7] Mayo Foundation for Medical Education and Research. (2021, July 23). Heart scan (coronary calcium scan). Mayo Clinic. Retrieved May 12, 2022, from https://www.mayoclinic.org/tests-procedures/heart-scan/about/pac-20384686

[8] Tackling missing value in dataset. Analytics Vidhya. (2021, November 12). Retrieved May 12, 2022, from https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value

[9] https://github.com/SahuH/CS760-project