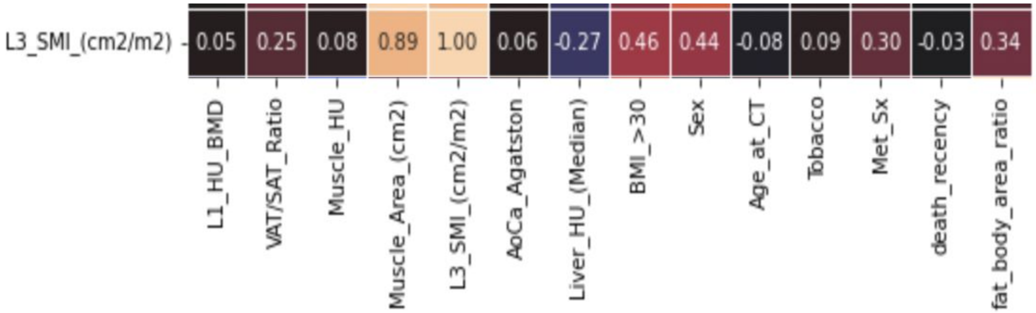
Estimating adverse clinical outcomes and Biological Age using CT & Clinical Data

Harsh Sahu
Ganesh Cheerla
Hemal kumar Patel

Preprocessing and Feature Engineering

- Clipping of higher values to remove skewness in few features.
  Example : AoCa_Agatston clipped at 99 percentile

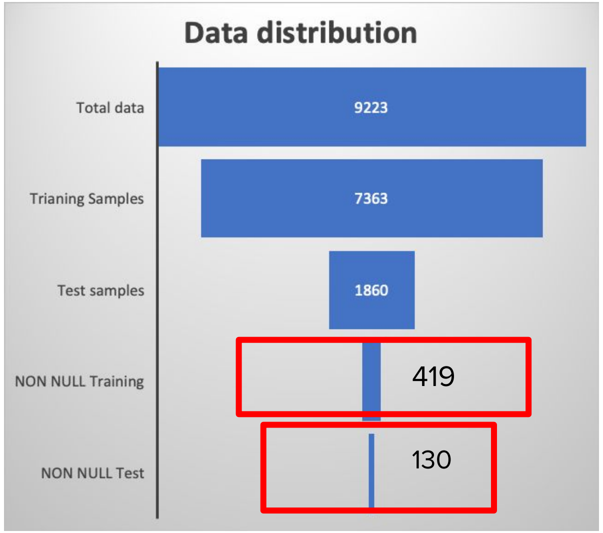
- Fill NULL values in few CT features using **iterative imputing** based on other features.
  - Example: L3_SMI_(cm2/m2) filled using 'BMI_more_than_30' and 'Sex'
  - Remaining filled with Median/Mean.



- Dropped some features based on correlation and created new features like "TAT/Body area."

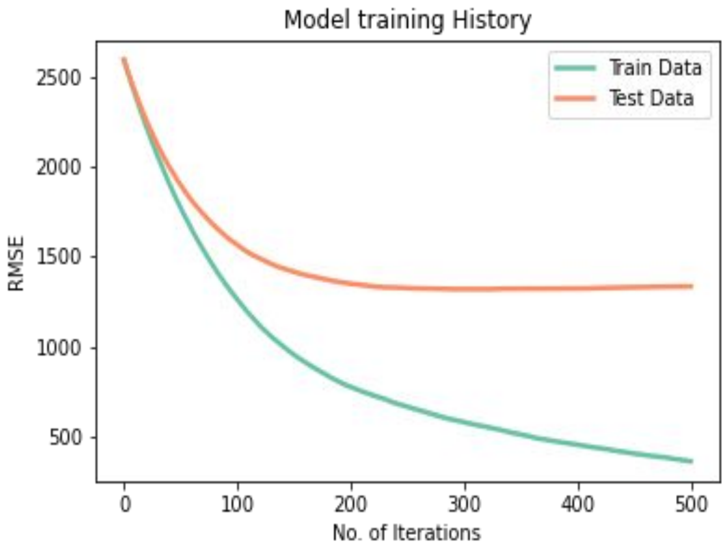
[Regression] Predicting No. of Death Days

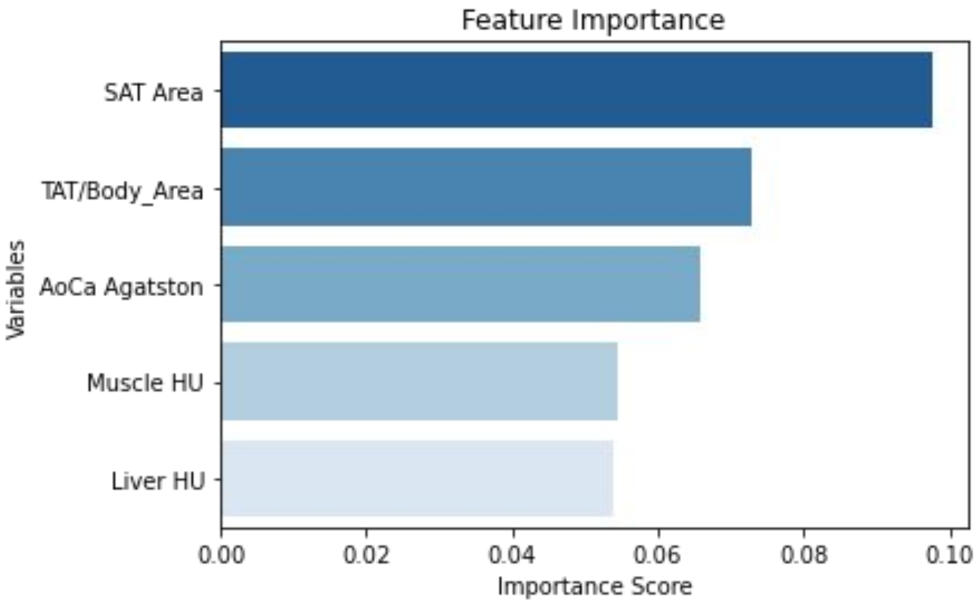
Sub-sampled the Data to people who have died (samples for which we have non-null values)



Model	Only CT (RMSE)	CT+Clinical (RMSE)
Linear Regression	1351	1324 (-2%)
SVR (Support Vector Regressor)	1410	1381 (-2%)
XGBoost	1331	**1314 (-1.2%)**

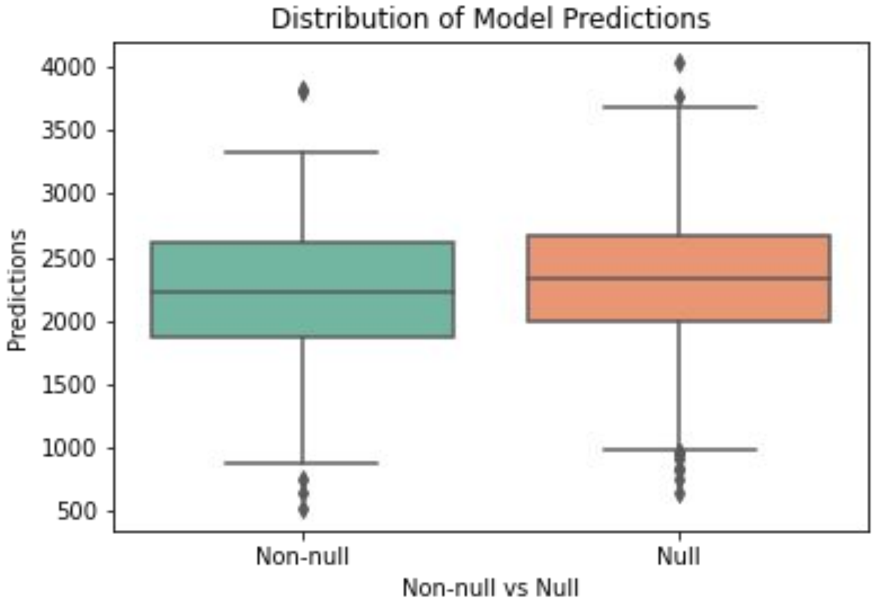
Best Model Results

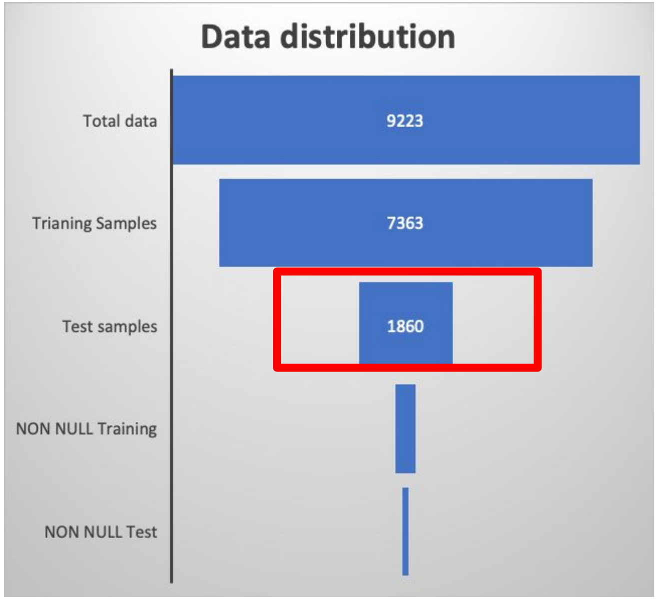




Prediction on NULL values

Let's take the trained model and try predicting on all the Test samples (NULL + Non NULL)

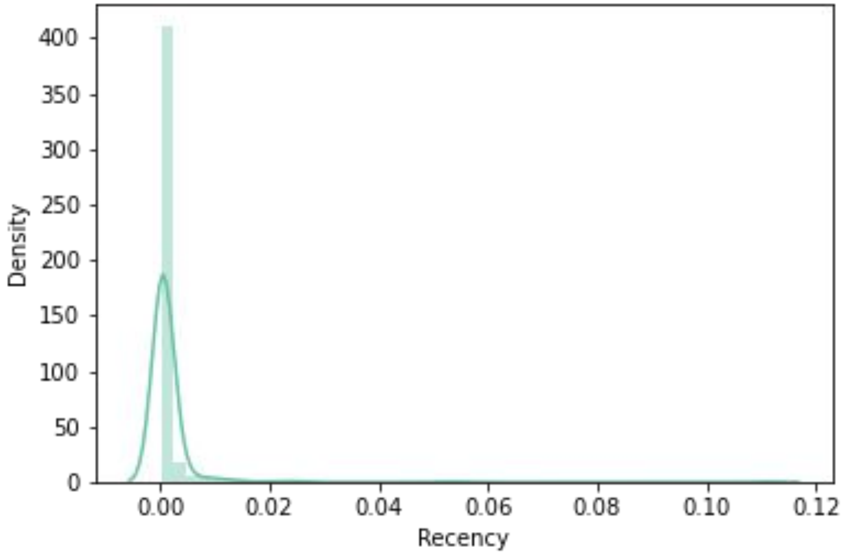




There is a need to somehow incorporate “Null” samples in Training

[Regression] Predicting “Recency”

- We define a new quantity “Recency”
- For people who have died: **Recency = 1 / No. of death days**
- For people who have not died: **Recency = 0**

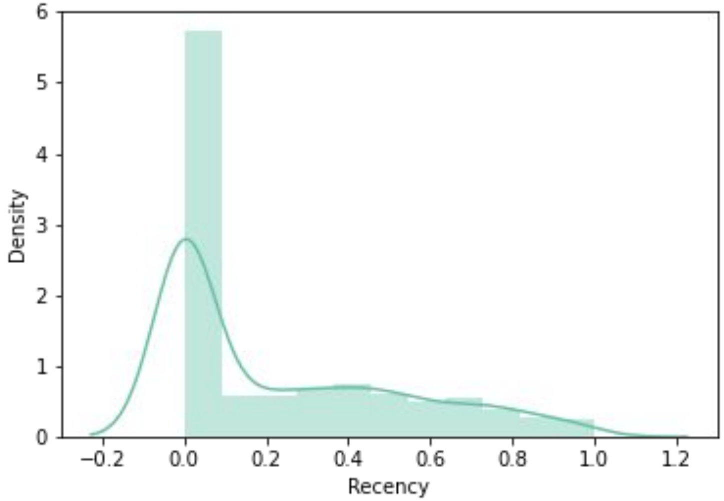


Highly Skewed :(

Incompatible to be Trained

Transforming Recency

- Transformed positive (>0) samples to a more uniform distribution using Box-Cox Transform
- Under-sampled “zero” samples to be comparable to non-zero samples



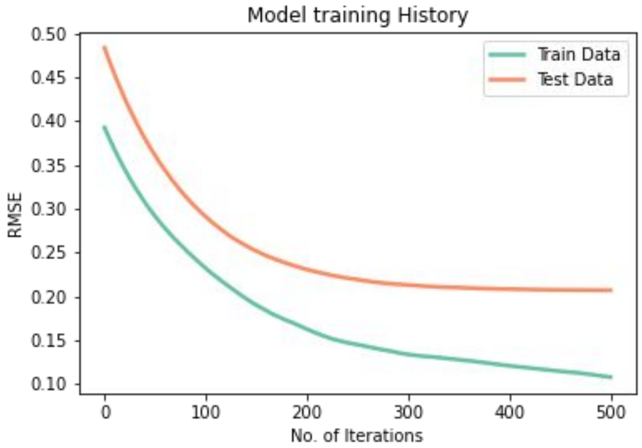
Final Training data = **838** samples

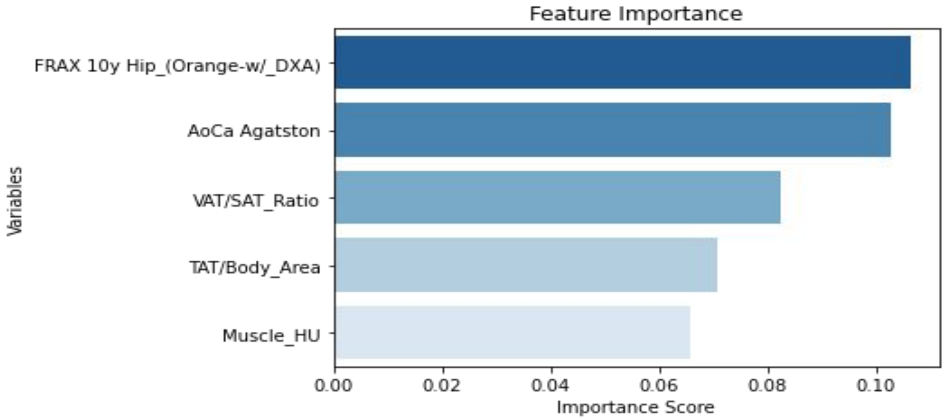
Test data = **1860** samples [We are going to predict for everyone :)]

Results

Model	Only CT (RMSE)	CT+Clinical (RMSE)
Linear Regression	0.235	0.212 (-9.7%)
SVR (Support Vector Regressor)	0.235	0.243 (+3.4%)
XGBoost	0.232	**0.204 (-13.36%)**

Analysing Best Model...



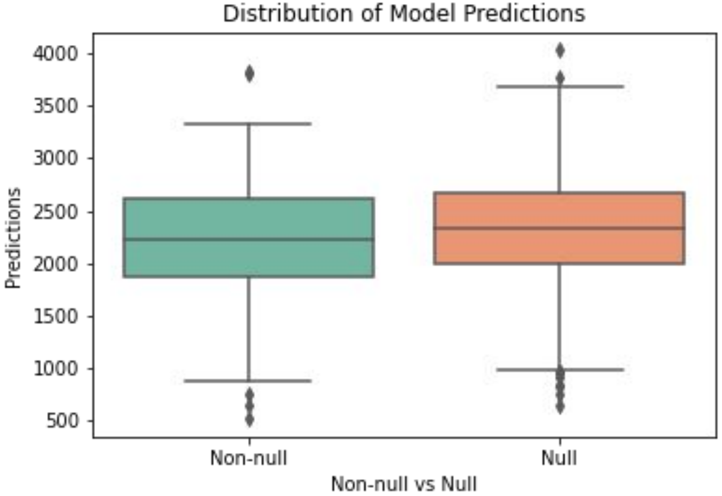


Predicting “Days” vs Predicting “Recency”

“Days”

(Using the Best Model)

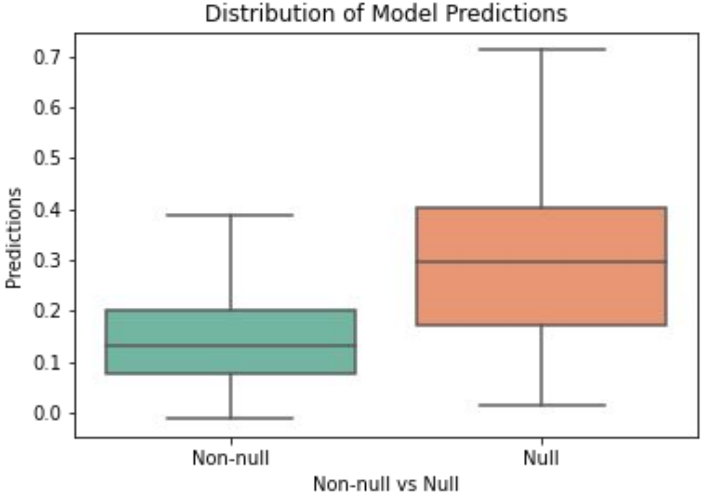
Error% = RMSE/(True values’ mean) = **58.85%**



“Recency”

(Using the Best Model)

Error% = RMSE/(True values’ mean) = **61.42%**



Predicting other clinical outcomes (using Recency)

Heart Attack

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

Aortic Calcification comes out to be the best predictor

Diabetes

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

Metabolic Syndrome comes out to be the best predictor

Biological Age

Methodology

Assumption:

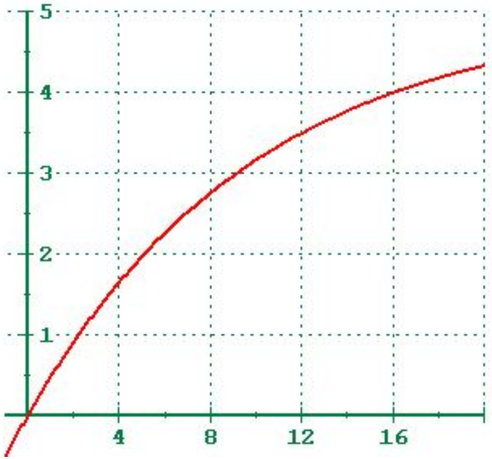
- People die at bio age 100
- Higher the “DEATH[d from CT]”, higher the patient has bio-days left

Data Processing:

- Split Data in train/test fashion using key column “DEATH [d from CT]” value. If non-empty -> train, else->test

Methodology:

- Train data: compute bio_days_left using the exponential decay increasing function shown in the right
- **Bio_age** = **max_bio_age - bio_days_left**
- Apply linear regression and XGBoost(Better)

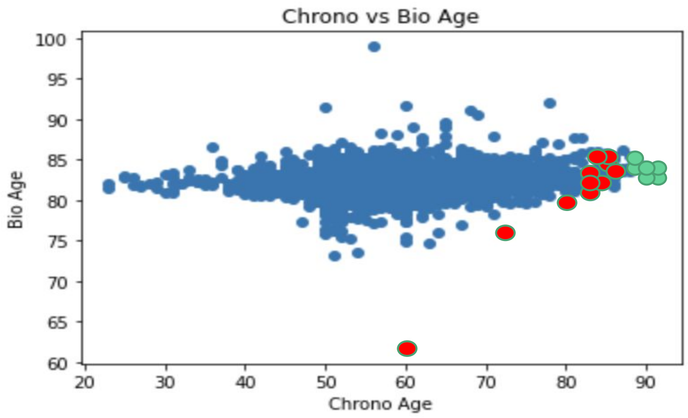


Sample: *bio_days_left using exponential decay increasing function*

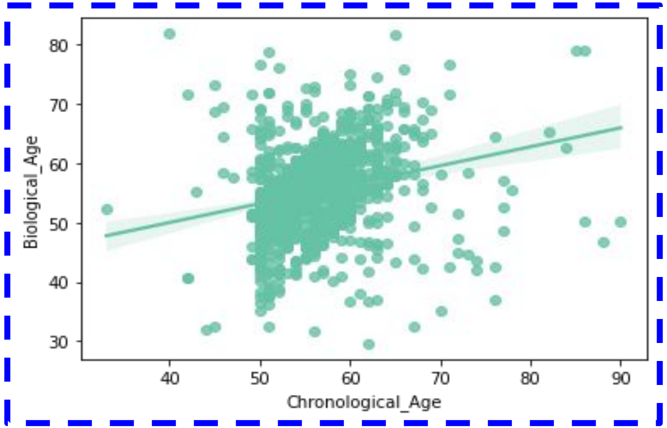
$$y = a + C \cdot \left(1 - e^{-kx}\right)$$

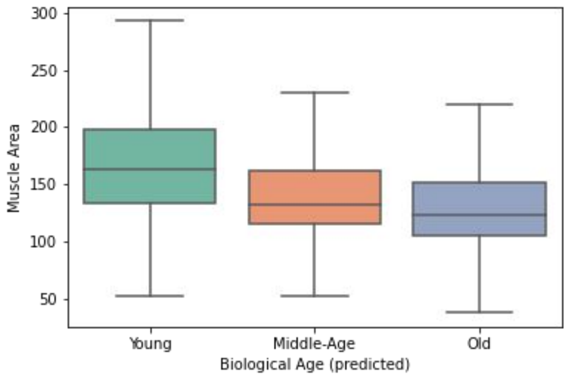
Result and Verification

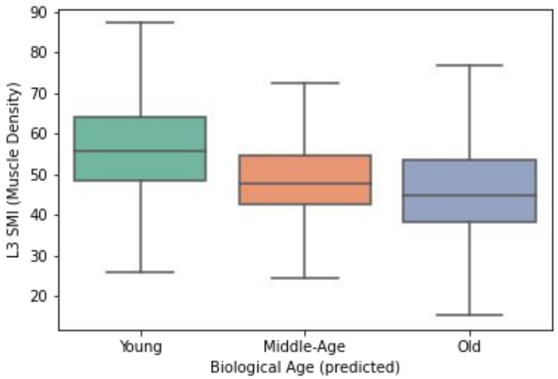
Linear Regression(doesn't work well)

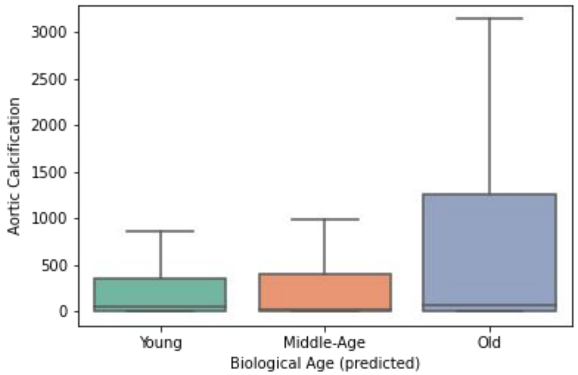


XGBoost (better)









Thank You!