Predicting Age at Death for Americans: Based on CDC's 2015 Mortality Dataset

Sibo Zhou
Data Science Institute, Brown University
Dec. 15, 2024
Github Repository: https://github.com/sibozhou99/data1030project.git

Section 1: Introduction

Motivation:
Life expectancy in the United States varies significantly across different demographic groups, influenced by factors such as sex, race, marital status, and ethnicity. Understanding the relationships between these demographic factors and age at death enables researchers, particularly in social sciences and public health, to identify issues like gender and racial inequalities for further investigation.

About the CDC Mortality Dataset:
Each year, the CDC publishes the most comprehensive report on deaths in the United States through the National Vital Statistics Systems [1], one of the oldest inter-governmental public health data-sharing mechanisms [5]. This system accesses vital registration data from government-operated birth and death records. The CDC mortality dataset we used records every death in the country from 2005 to 2015, including detailed information on causes of death, underlying illnesses, and demographic details of each death incident [1].

Previous Works (Kaggle projects, published papers):
Most existing Kaggle projects on the CDC Mortality dataset involve descriptive analyses, such as grouping age at death by gender or race. We have identified two projects closely related to my objective of exploring the relationship between mortality rates and demographic features:

"Who'll Die Because of Accidents?" This project uses demographic features to predict accidents (0 or 1) as a cause of death. It employs Logistic Regression (Accuracy: 63%), Random Forest (Accuracy: 65%), and XGBoost (Accuracy: 65%) in its ML pipeline. Compared to a baseline accuracy of 57%, these models perform better, though results remain modest. Resident_status and race emerge as significant features [2].

"Predicting Workplace Injuries" This project targets injury_at_work (0 or 1) as the cause of death, framing it as a classification problem. It utilizes Logistic Regression and XGBoost, with marital_status and sex identified as the most important features [3].

Additionally, numerous peer-reviewed articles analyze the CDC mortality data to explore the relationship between demographic factors and mortality. For example, a recent public health paper by Holmes Jr. et al. [4] uses statistical methods to examine disproportionate mortality trends among black children. The study concludes that racial disparities in infant mortality, particularly among Black populations, can be partly attributed to higher incidences of medical misadventures and traffic accidents, confirming correlations between demographic features like race and age at death [4].

Section 2: Exploratory Data Analysis

We examined basic information about each demographic variable and the target variable, detail_age. To explore relationships further, we created a distribution plot for detail_age and used box plots, violin plots, and category-specific histograms to visualize its interactions with key demographic variables.

The dataset consists of 2.7 million rows and 77 columns, making it quite large. For the project, we narrowed our focus to relevant demographic features: resident_status, marital_status, sex, race, hispanic_origin, education.

Target Variable (detail_age):

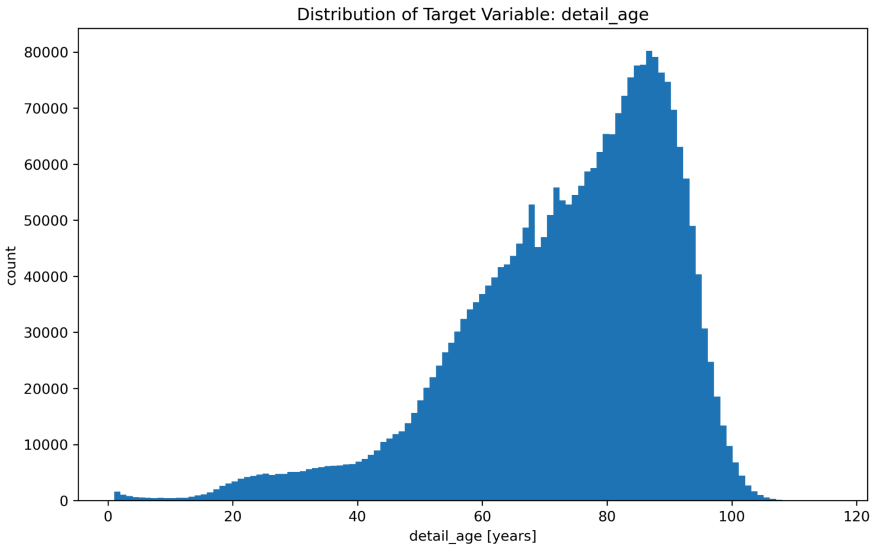

Figure 1. Distribution of target variable: detail_age. X-axis is detail_age in years and y-axis is the count.

Our target variable, detail_age, represents the age at death (in years) for each record in the dataset. The descriptive statistics are:

Median: 77 years
Maximum: 116 years
Minimum: 1 year
SD: 18.08 years

The distribution plot reveals that detail_age is right-skewed. Key observations include:

80% of the data has detail_age over 60 years.
Less than 5% of the data has detail_age below 40 years.

This significant imbalance in the target variable will be addressed during data preprocessing and splitting to ensure accurate model training and evaluation.

Detail_age and marital_status

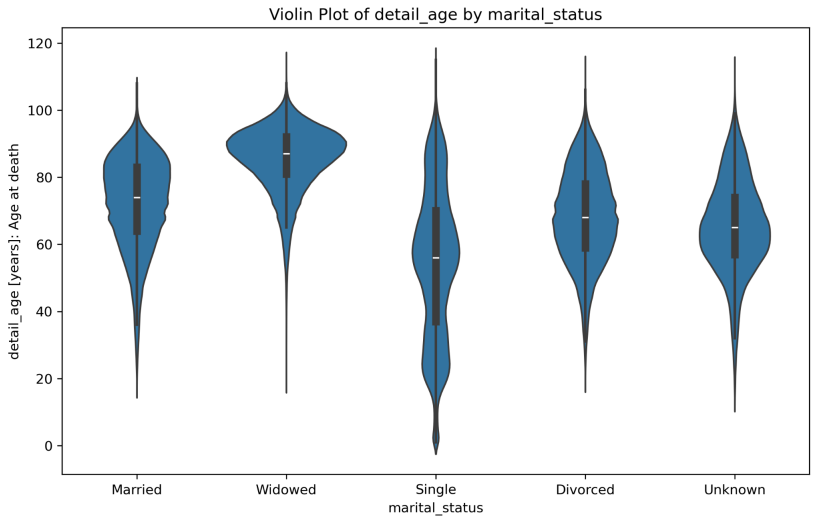

Figure 2. Violin plot of detail_age by marital_status. X-axis are different marital statuses. Y-axis is detail_age in years.

The relationship between detail_age and marital_status reveals significant disparities:

Widowed: Median age at death is 87 years.
Married: 74 years.
Divorced: 68 years.
Single: 56 years.

These differences suggest that marital_status is likely to be a highly predictive variable in the

ML models.

Detail_age and race

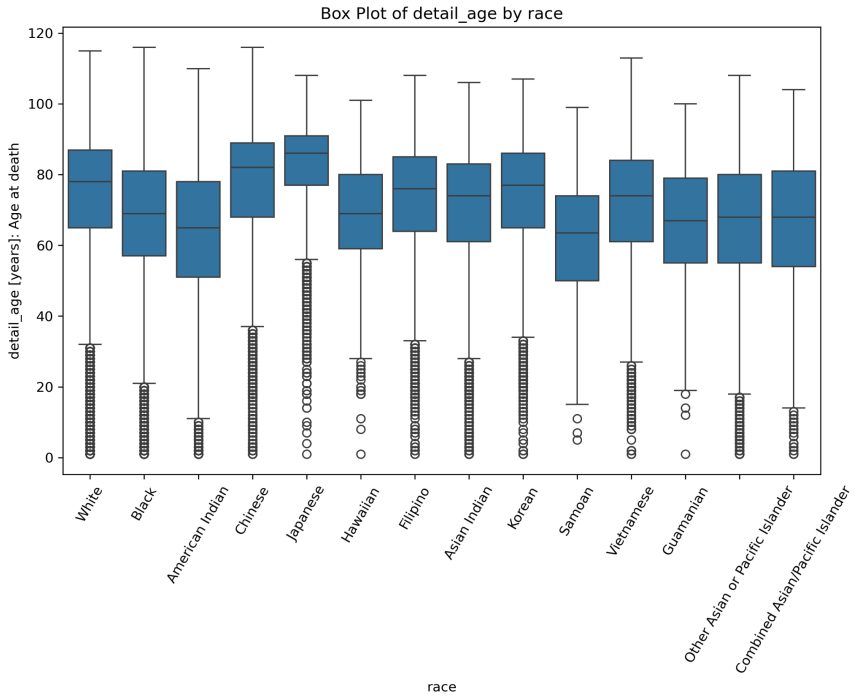

Figure 3. Box plot of detail_age by race. X-axis are different races. Y-axis is detail_age in years.

There are significant disparities in detail_age across racial groups:

Japanese: Median age at death is 86 years.
White: 78 years.
Black: 69 years.
Samoan: 63.5 years.

These differences highlight that underprivileged racial groups, such as African Americans and Pacific Islanders, tend to have substantially lower median ages at death compared to White or East Asian groups.

Detail_age and sex

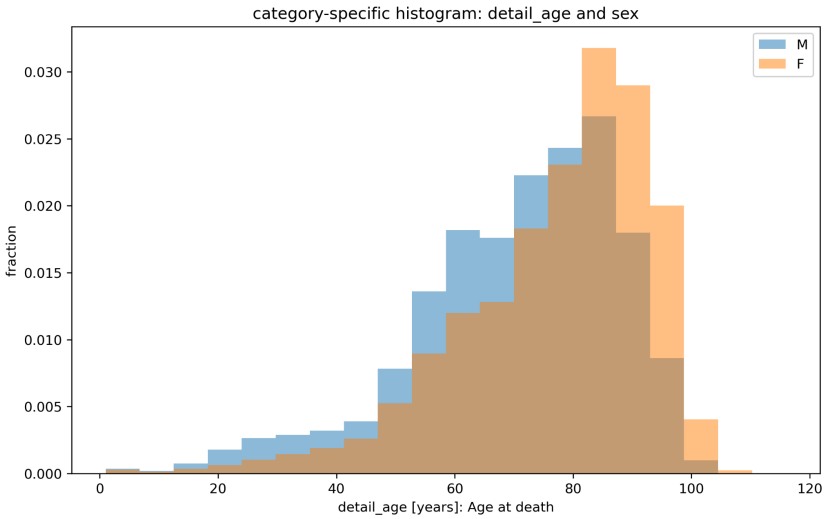

Figure 4. Category-specific histogram of detail_age and sex. Yellow is female, blue is male. X-axis is the detail_age in years and the y-axis is the fraction of data points (normalized frequency) for each detail_age bin.

The median age at death is 81 years for females and 73 years for males. The histogram reveals that females have a higher proportion of data points at older ages, indicating they are overrepresented in the upper ranges of detail_age.

Section 3: Methods

Data Cleaning and Preparation

We removed 0.02% of rows where detail_age was 999 ("age not stated") and created a new variable, hispanic_category, by mapping hispanic_origin to CDC-defined categories. Missing values in education_2003_revision were filled using data from education_1989_revision, eliminating all nulls in the education variables. Finally, we selected seven key columns, including demographic variables and the target variable, for the next step in the pipeline.

ML Models	**Hyperparameters**	**Values (Best Model in Red)**
Lasso (Linear Regression with L1 Regularization)	alpha	[0.001, 0.01, 0.1, 1, 10, 100]
Ridge (Linear Regression with L2	alpha	[0.001, 0.01, 0.1, 1, 10, 100]

Regularization)		
Elastic Net (Linear Regression with L1+L2 Regularization)	alpha l1_ratio	[0.001, 0.01, 0.1, 1, 10, 100] [0.01, 0.25, 0.5, 0.75, 1]
RandomForestRegressor	max_depth max_features	[1, 5, 9, 12, 20, 50] [0.1, 0.325, 0.55, 0.775, 1.0]
XGBRegressor (With Early Stopping = 30)	learning_rate subsample max_depth	[0.03] [0.6, 0.8] [10, 20, 30, 50]

Table 1. Hyperparameters of ML models (hyperparameters of the best models are in red)

ML Pipeline Structure Overview

We evaluated model stability and robustness by looping through 5 random states, performing the following steps within each iteration:

Splitting Strategy
To address the right-skewness of the target variable detail_age (as shown in the EDA), we used StratifiedShuffleSplit to divide the data into 20% testing and 80% training/evaluation. Stratification by five detail_age bins (<20, 20–40, 40–60, 60–80, >80) ensured proportional representation of age groups and prevented sampling bias.

Preprocessing Pipeline
OneHotEncoder: Encoded categorical variables (sex, marital_status, race, hispanic_category, resident_status) into binary columns.
OrdinalEncoder: Encoded education levels in their natural order.
StandardScaler: Standardized features to have a mean of 0 and standard deviation of 1.

Preprocessing increased the number of independent variables from 6 to 38.

Hyperparameter Tuning
We used GridSearchCV with 3-fold CV for Lasso, Ridge, Elastic Net, and RandomForestRegressor, combining preprocessing, scaling, and modeling into a single pipeline.

RMSE was the scoring metric for optimization. For XGBoost, since GridSearchCV does not support early stopping, we manually performed 3-fold CV to find the best hyperparameters (Table 1, best values highlighted in red).

Evaluation Metrics
Baseline RMSEs, calculated using the mean age at death, served as benchmarks. Test RMSEs, computed on the 20% test set, measured final model performance. RMSE was chosen for its interpretability (same unit as the target variable) and sensitivity to large errors, critical for predicting age_at_death.

At the end of each loop, we saved the best model, hyperparameters, test RMSE, baseline RMSE, and test data (x-test, y-test).

To measure splitting uncertainty, we calculated the standard deviation of test scores over 5 random states for Lasso, Ridge, and Elastic Net, which is 0.0102. For non-deterministic model uncertainties, we calculated the S.D. of test scores for RandomForestRegressor (0.0116) and XGBoostRegressor (0.0109).

Section 4: Results

ML Model	Mean Test RMSE	S.D. of Test RMSE	Mean Baseline RMSE	(Mean Test RMSE - Mean Baseline RMSE) / Standard Deviation
Lasso	13.67	0.0102	17.07	333
Ridge	13.67	0.0102	17.07	333
Elastic Net	13.67	0.0102	17.07	333
RandomForest Regressor	13.24	0.0116	17.07	330
XGBRegressor	13.25	0.0109	17.07	350

Table 2. Test RMSE results and baseline RMSE results of ML models. RandomForestRegressor in red achieves the lowest RMSE among all models.

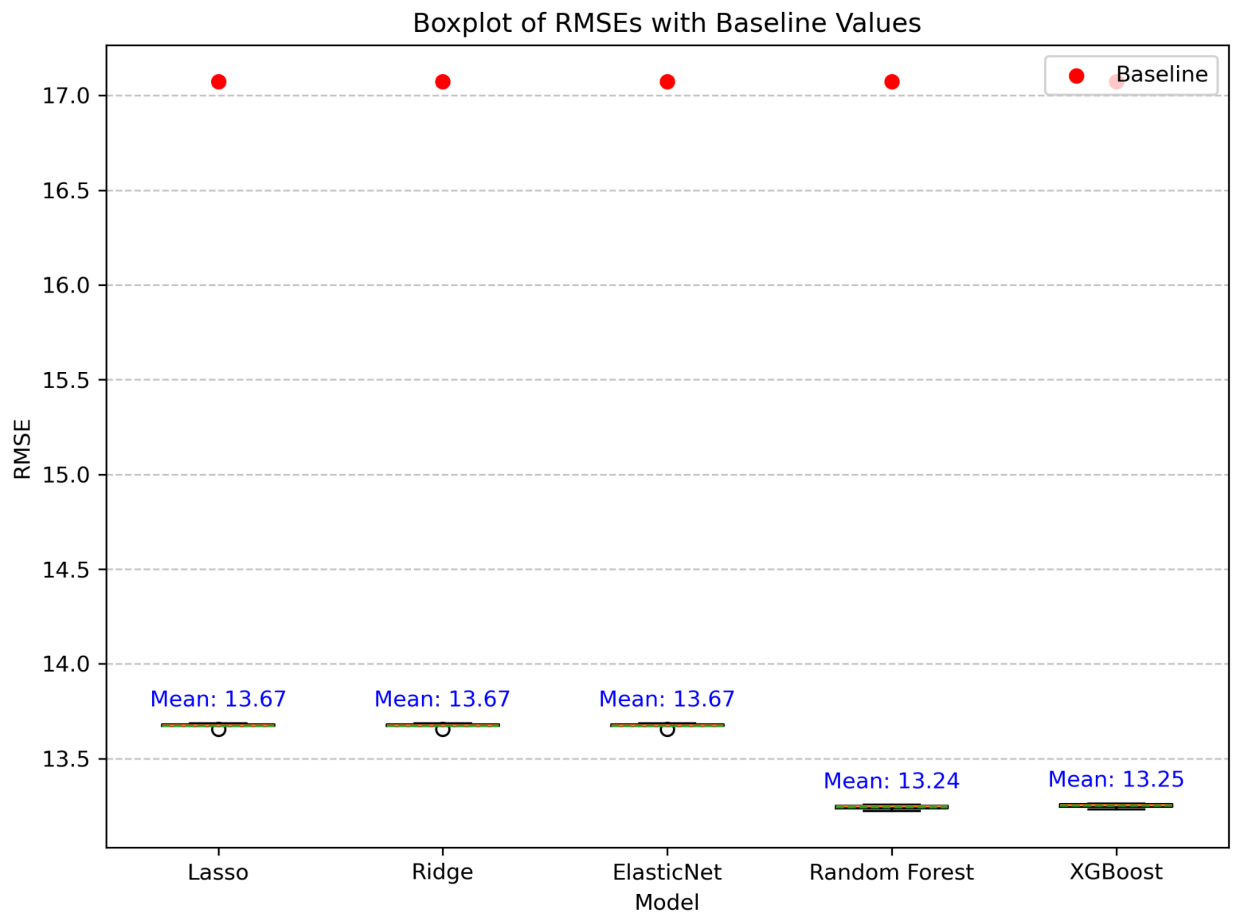

Figure 5. Box plot of model test RMSEs with baseline RMSEs. X-axis are the five ML models, the y-axis is the RMSE. Red dots represent the baseline RMSEs.

All ML models surpassed the baseline RMSE of 17.07. RandomForestRegressor achieved the lowest test RMSE at 13.24, outperforming XGBRegressor at 13.25. Linear models—Lasso, Ridge, and Elastic Net—each had an RMSE of 13.67. Normalized by standard deviation, XGBRegressor demonstrated the greatest improvement (350 above baseline), followed closely by RandomForestRegressor (330). RandomForestRegressor is most predictive, with tree-based models outperforming linear models overall.

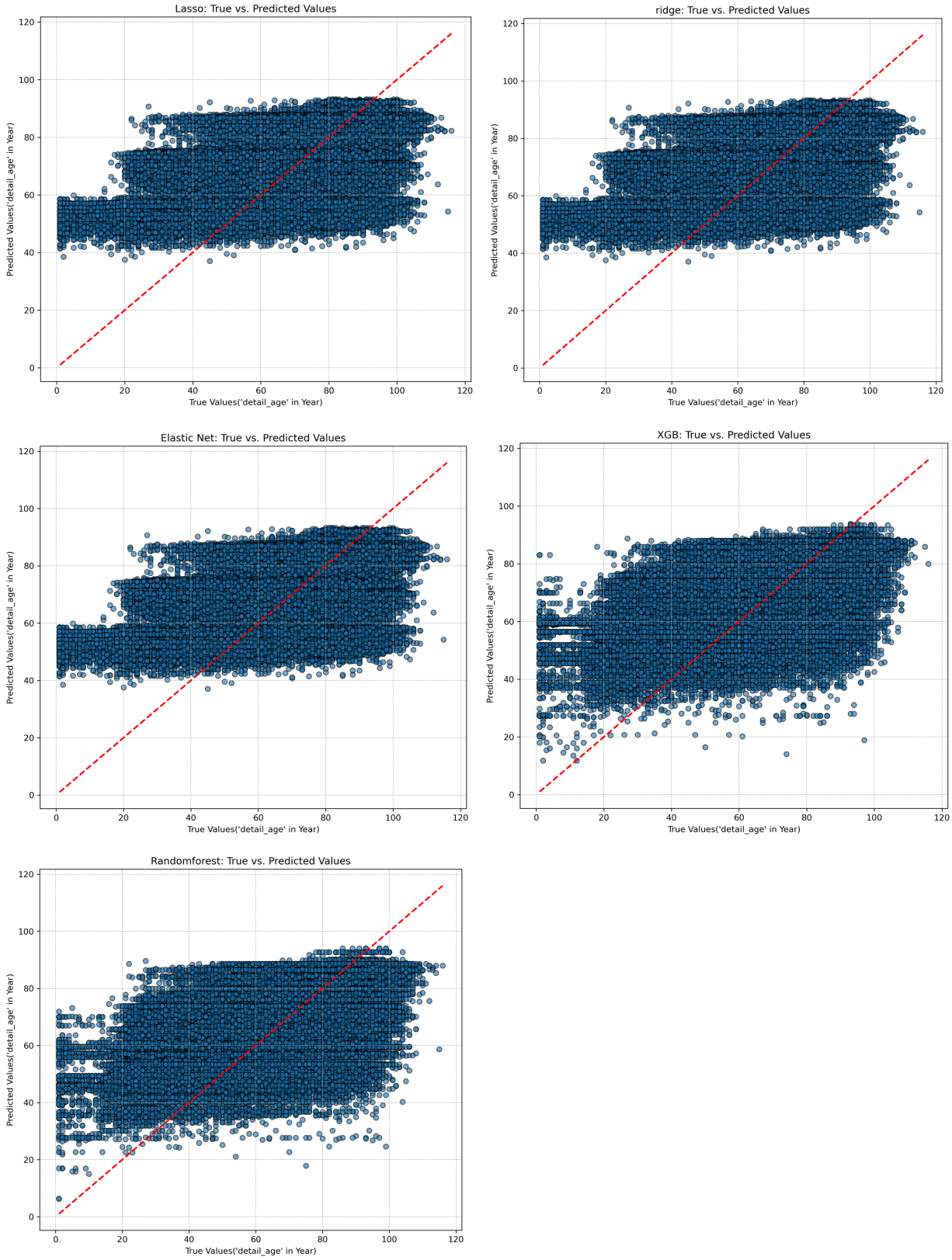

Figure 6. Scatter plots of true value vs predicted value of target variable ‘detail_age’ for all five ML models. The red diagonal dash line represents the perfect prediction.

We used scatter plots to compare the true values of detail_age with model predictions. Predictions across all models were more dispersed than expected, deviating from the perfect prediction line (red diagonal dash line). Non-linear models like RandomForestRegressor and XGBRegressor performed slightly better, with predictions closer to the diagonal.

Global Feature Importance

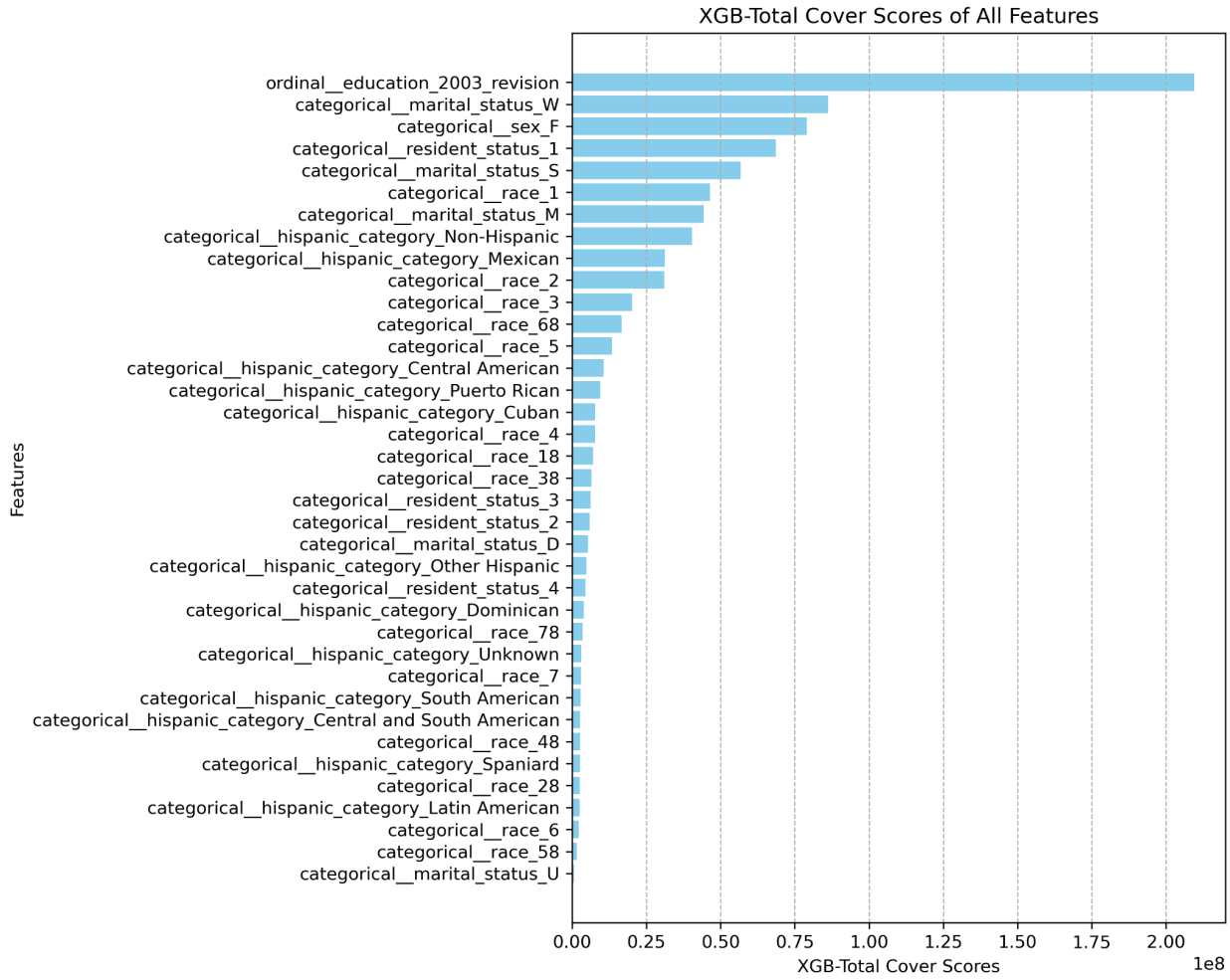

Figure 7. XGB-Total cover scores of all the features, indicating global importance.

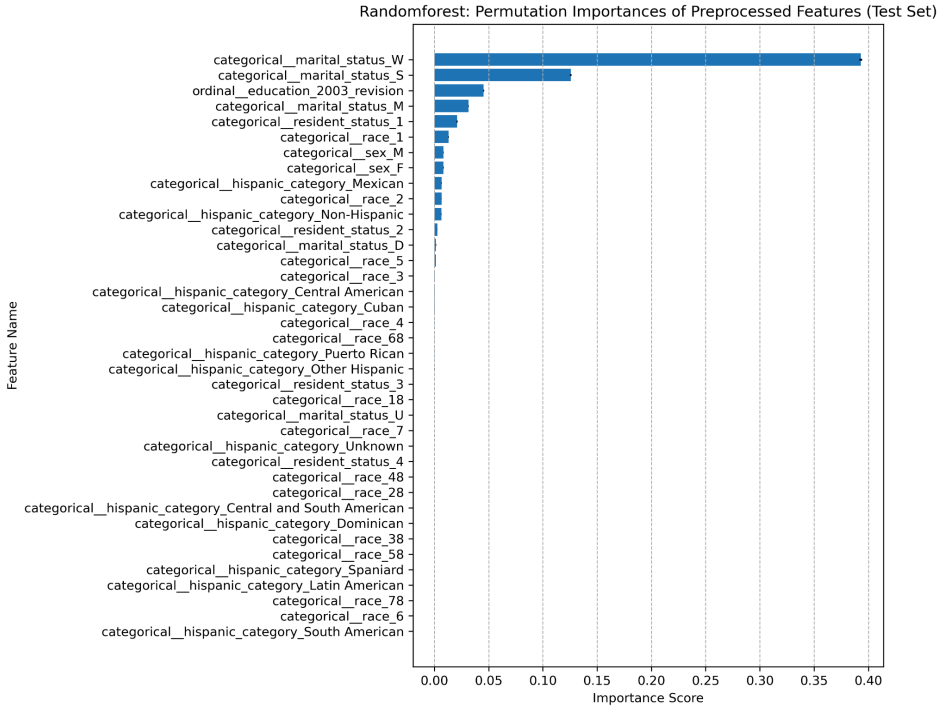

Figure 8. RandomForest–permutation importance scores of all the features, indicating global importance.

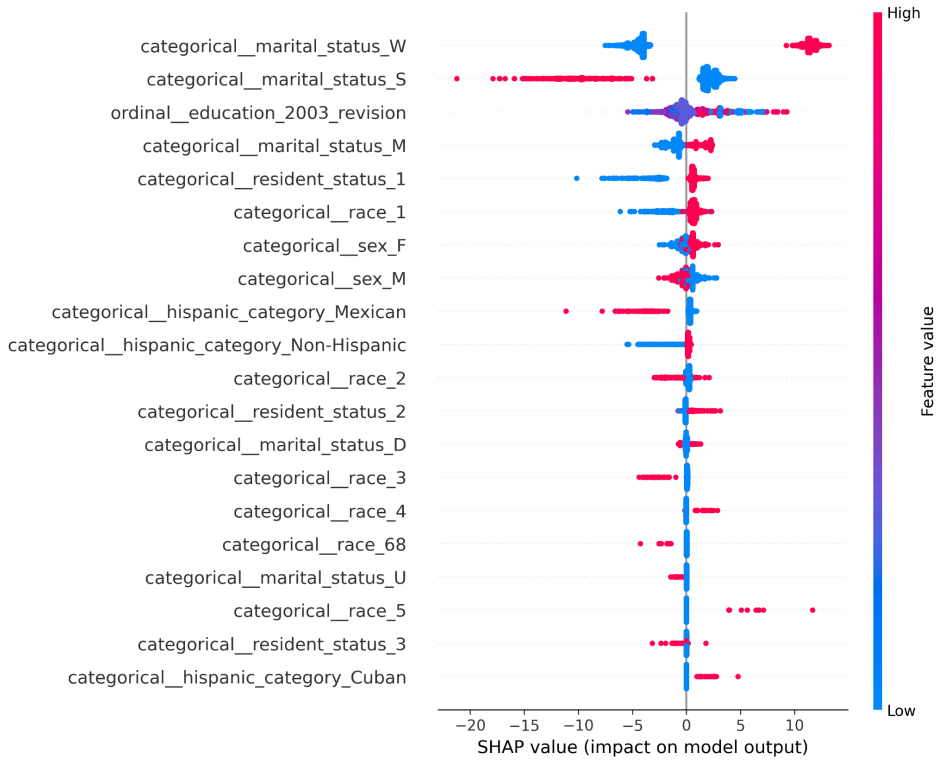

Figure 9. RandomForest–SHAP summary plot of all the features, indicating global importance.

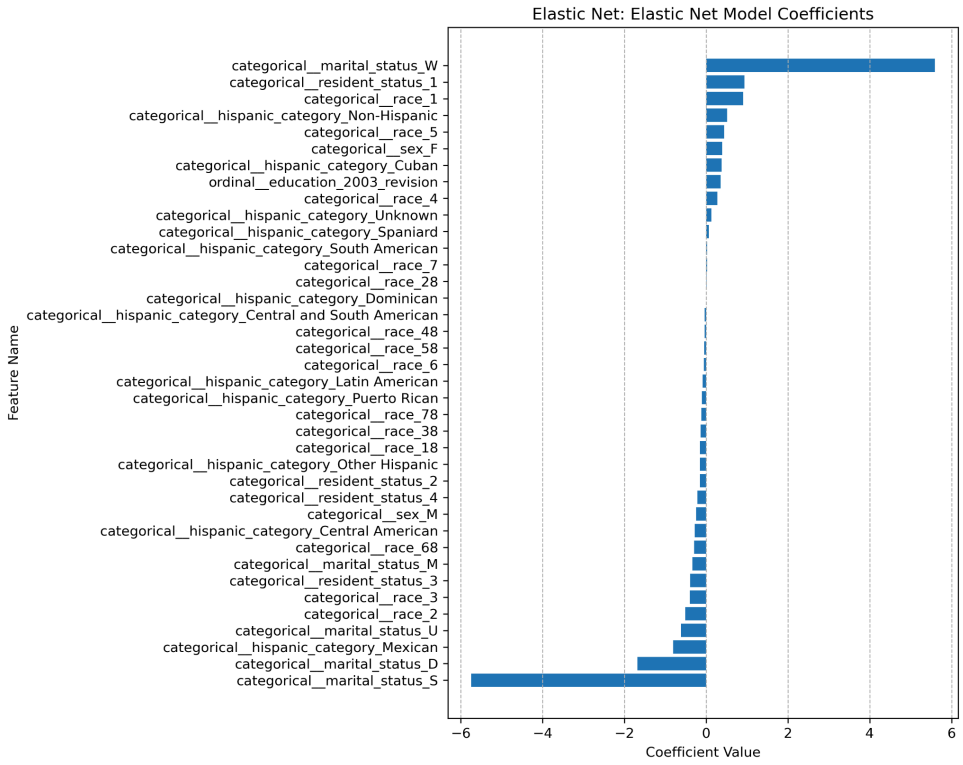

Figure 10. Elastic Net–coefficients of all the features, indicating global importance.

We analyzed the global importance of features for predicting age_at_death, as shown in Figures 7–10, using XGBoost (total cover scores), RandomForest (permutation importance and SHAP summary), and Elastic Net (feature coefficients). Permutation importance (Figure 8) showed that shuffling marital_status caused the largest drop in test score, highlighting its significance. Elastic Net coefficients (Figure 10) further revealed that being widowed or a local resident positively impacted predictions, while being divorced or single had a negative effect. Consistently across all models, marital_status, education level, and sex emerged as the most predictive features.

Local Feature Importance

Figure 11. XGB–SHAP force plot of index = 20, with predicted value lower than base value, indicating local importance.

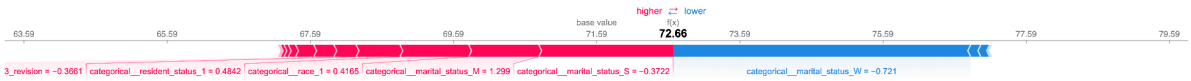

Figure 12. XGB–SHAP force plot of index = 40, with predicted value higher than base value, indicating local importance.

For the data point at index 20, the SHAP force plot for the XGBRegressor reveals the local importances of the features. Being “not single” (marital_status_S = 0 before scaling), “white”, “a local resident” increased the prediction. Conversely, being “not widowed” decreased the prediction. Overall predicted age at death is 70.81 years, lower than the base value of 71.59 years.

For the data point at index 40, being “not single”, “married”, “white”, and “a local resident” increases the prediction. At the same time, being “not widowed” decreases the prediction. Overall predicted age at death is 72.66 years, higher than the base value of 71.59 years, reflecting the combined influence of these features.

Section 5: Outlook

To further improve the model and its interpretability, we can consider using instrumental variables for better causal inference: While marital_status is the most globally important variable, its strong correlation with age_at_death raises concerns about reverse causality, where the outcome may influence the predictor. Using effective IVs helps to generate more reliable interpretations of the relationship between marital status and age at death.

From an accuracy standpoint, the inclusion of non-demographic variables, such as health conditions or data on chronic diseases, could enhance overall model performance. These factors explain a much larger portion of the variability in age_at_death than demographic variables. Additionally, exploring advanced feature engineering techniques could improve predictive power. For instance, creating interaction terms like marital status × sex could capture nuanced relationships between variables that are otherwise overlooked in the current model. Finally, we

can experiment with deep-learning models such as neural networks that could help capture the intricate, non-linear relationships in the data.

References

[1] Centers for Disease Control and Prevention (CDC). "Mortality Data." Kaggle, https://www.kaggle.com/datasets/cdc/mortality/data.

[2] Ran. "Who'll Die Because of Accidents?" Kaggle, https://www.kaggle.com/code/changran/who-ll-die-because-of-accidents.

[3] DI. "What Kind of People Will Get Injured at Work?" Kaggle, https://www.kaggle.com/code/jswxhd/what-kind-of-people-will-get-injury-at-work/notebook.

[4] Holmes, Jr, et al. “Medical Misadventures as Errors and Mistakes and Motor Vehicular Accidents in the Disproportionate Burden of Childhood Mortality among Blacks/African Americans in the United States: CDC Dataset, 1968-2015.” *Healthcare (Basel)*, vol. 12, no. 4, 2024, pp. 477-, https://doi.org/10.3390/healthcare12040477.

[5] Centers for Disease Control and Prevention (CDC). "About the National Vital Statistics System." Centers for Disease Control and Prevention, https://www.cdc.gov/nchs/nvss/about_nvss.htm.