Mortality: CS109a Final Project

Fall 2020

Leonard Tang, Ray Chen, Vineet Gangireddy, and Sean Ty

Contents

1 Problem Statement 2

2 Pre-Processing the Data 2

3 Feature Descriptions 3

4 Understanding the Data (Selected EDA) 3

5 Methods and Interpretations 4

5.1 K-Nearest Neighbors 4

5.2 Logistic Regression 4

5.2.1 Logistic Regression with Lasso regularization 5

5.2.2 Logistic Regression with Ridge regularization 5

5.2.3 Interpretation from Logistic Regression Models 5

5.3 Decision Trees 5

5.3.1 Single Tree 5

5.3.2 Random Forest 7

5.3.3 Bagging 7

5.3.4 Boosting 9

6 Conclusion 12

7 Impact Statement 14

8 Appendix 15

8.1 SHAP Theory 15

8.2 Full-Scale EDA Plots 16

8.3 K-Nearest Neighbors Plots 17

1 Problem Statement

Given the dataset from the NHANES I Epidemiologic Follow-up Study, a longitudinal study analyzing clinical, nutritional, and behavioral factors and their relationship with subsequent morbidity, mortality, and hospital utilization, we set out to build a model predicting mortality within 18 years of the initial study by comparing a variety of k-NN models, logistic regression models, and decision tree-based models. We implemented ensemble methods, such as bagging, boosting, and random forests, on top of our decision trees to maximize our ROC AUC scores. We also briefly investigated similarly grouped features, such as demographic data, serum levels, and iron-related features (e.g. serum iron, TS, TIBC), to see if we could find a “minimally expressive" feature set consisting of the most informative predictors. ***We focused our efforts on interpreting and understanding our models, more so than purely maximizing model performance.*** To that end, we analyzed feature importance, permutation feature importance, and SHAP values to discern the most informative features for predicting mortality.

2 Pre-Processing the Data

To clean the data, we first dropped the indexing column named “Unnamed: 0"; checked data types in the columns to see if any needed to be casted; and took care of NaN/missing values by counting how many rows there were with at least one NaN value, counting how many NaN values each column/feature had, and examining rows with at least one NaN value, ultimately imputing with a k-NN model with 1 neighbor. We also reformatted our response variable from being a quantitative variable denoting the number of years between death and initial examination to a ***categorical variable indicating survival (labelled 1) of at least 18 years following initial examination***. According to the dataset, a negative response variable $-y_i$ indicated a patient survived for at least y_i years; by setting a classification threshold of 18 years, we were able to partition the data into a roughly even 50/50 class split. For EDA, we constructed correlation matrices for the predictor and response variables; constructed histograms for each predictor variable (to check for outliers); and ran data.describe() to compare subsets of the data. Finally, we added a baseline model (a simple k-NN model) to get a lower bound on model performance. This yielded an ROC AUC score of 0.5.

3 Feature Descriptions

The dataset contains the following variables: **Age** (age of survey participant at time of the study); **Diastolic BP** (normal range 60-80 mmHg); **Poverty Index** (as measured by the Census bureau, where higher values indicate less poverty); **Red Blood Cells** (hundred thousands per microliter); **Sedimentation Rate** (measuring bodily inflammation); **Serum Albumin** (measure of albumin protein in the blood); **Serum Cholesterol** (measure of total cholesterol in the blood); **Serum Iron** (measure of blood iron); **Serum Magnesium** (measure of magnesium in blood); **Serum Protein** (measure of total protein in blood); **Sex** (with 0 = Male and 1 = Female); **Systolic BP**; **TIBC** (Total Iron Bonding Capacity); **TS** (Transferrin Saturation, measuring plasma quantity in blood); **White Blood Cells** (thousands per microliter); **BMI**; **Pulse Pressure** (Systolic BP - Diastolic BP); **Other** (1 indicates race is not white or black); and **Black** (1 indicates race is Black).

4 Understanding the Data (Selected EDA)

First, we obtained the correlation matrices for the predictor variables as well as the predictor variables *and* response variables belonging to the dataframe with dropped NaN values and the dataframe with imputed values:

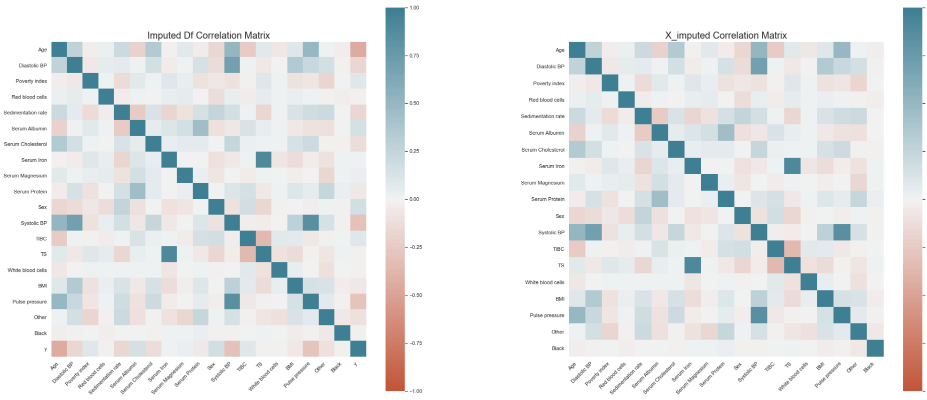

Figure 1: Correlation Matrices of Predictor Variables and All Variables

At first glance, it seemed that survivability was moderately correlated with age, systolic BP, as well as poverty index. We also plotted the histogram and kernel density plot of raw response variable and predictors to get a better sense of the distribution of the data. To see the distributions more clearly, please refer to the appendix.

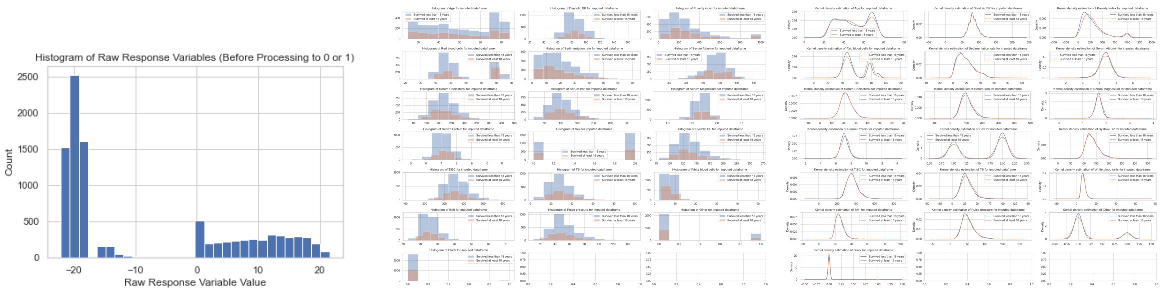

Figure 2: Distributions of the Predictors and Response Variable

In particular, we notice that the data is quite right-skewed, and that many participants exhibited an extreme negative response variable value, implying that most people had survived or were lost track of by the end of the study. Consequently, our classification problem was targeted towards predicting if an individual will survive at least 18 years.

5 Methods and Interpretations

5.1 K-Nearest Neighbors

The first model we decided to use was non-parametric, non-tree kNN classification. We used 5-fold cross validation to choose the best number of neighbors k and fit our final model to this best k value of 100 to find an AUC of 0.6848. Using this best-knn model, we found the permutation importance to see the most relevant predictors as given in Figure 8 of the appendix. This plot shows that following the process of computing permutation importances, the 6 most important features in descending order of importance for predicting survival for at least 18 years since the first examination were *Age, Sex, Pulse Pressure, Serum Albumin, Poverty Index, and Systolic BP.*

5.2 Logistic Regression

We first fit an unregularized logistic regression model as reference point, which yielded an ROC AUC score of 0.734. We created 4 additional logistical regression models: two logistic regression with Lasso regularization (one with and one without Cross-Validation) and two logistic regressions with Ridge regularization (one with and one without Cross-Validation). Across all the logistic regressions, we fed in the standardized imputed DataFrame.

5.2.1 Logistic Regression with Lasso regularization

We then fit a logistic regression with lasso regularization, where the inverse regularization strength was 1. The corresponding test AUC was 0.734. The predictors that had a positive relation with the response variable include: *Poverty index, Serum Albumin, and Sex.* Meanwhile, predictors that exhibited a negative relation with the response include: *Age, Systolic BP, and Pulse pressure.* Meanwhile Diastolic BP had no relation with the response. The Cross-Validation Logistic Regression with Lasso used inverse regularization strengths from 1e-4 to 1e4, and exhibited a similar accuracy and beta coefficients.

5.2.2 Logistic Regression with Ridge regularization

For ridge-like logistic regression with inverse regularization strength 1, the corresponding test accuracy was 0.734. The positive and negative predictors were similar to the logistic regression with Lasso. The Cross-Validation Logistic Regression with Lasso used inverse regularization strengths from 1e-4 to 1e4, and exhibited a similar accuracy and beta coefficient.

5.2.3 Interpretation from Logistic Regression Models

Across the 4 variations of logistic models, they all had similar ROC AUC scores to the unregularized logistic regression model, suggesting that regularization was not particularly effective. Likewise, the beta coefficients were similar across all 4 logistic regression (e.g. negative and large magnitude for Age, positive and large magnitude for Serum Albumin). The most influential predictors included Age, Sex, and Serum Albumin. However, across the different models there still exist differences: the Lasso regularized models suggest that there is no relation between Diastolic BP and log odds the response being 1, yet the Ridge models suggest otherwise.

5.3 Decision Trees

5.3.1 Single Tree

For the single decision tree model, we elected to perform 5-fold cross-validation to choose the best depth (maximizing the ROC AUC score) for our decision tree model. Ultimately, we found that the depth maximizing the validation set ROC AUC score was 4, yielding a score of 0.709. Below we plot the cross validation training results as well as ROC curve:

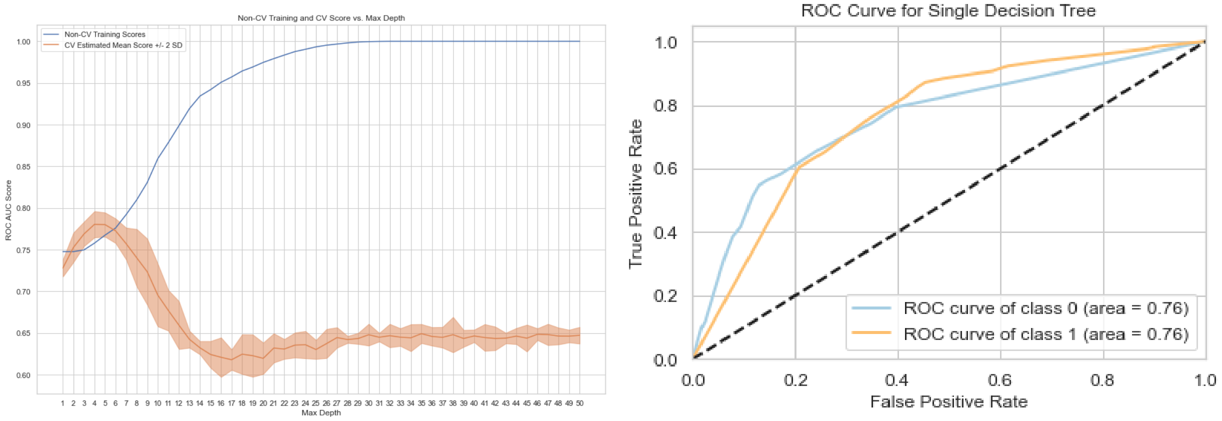

Figure 3: The Training Results and ROC Curve for a Single Decision Tree

To better understand and interpret our decision tree model, we elected to use SHAP (SHapley Additive exPlanations) values (please refer to Appendix 8.1 for a more thorough theoretical discussion), a model interpretabiltiy framework derived from the notion of Shapley values in game theory. We plot the mean SHAP values for our single decision tree model below:

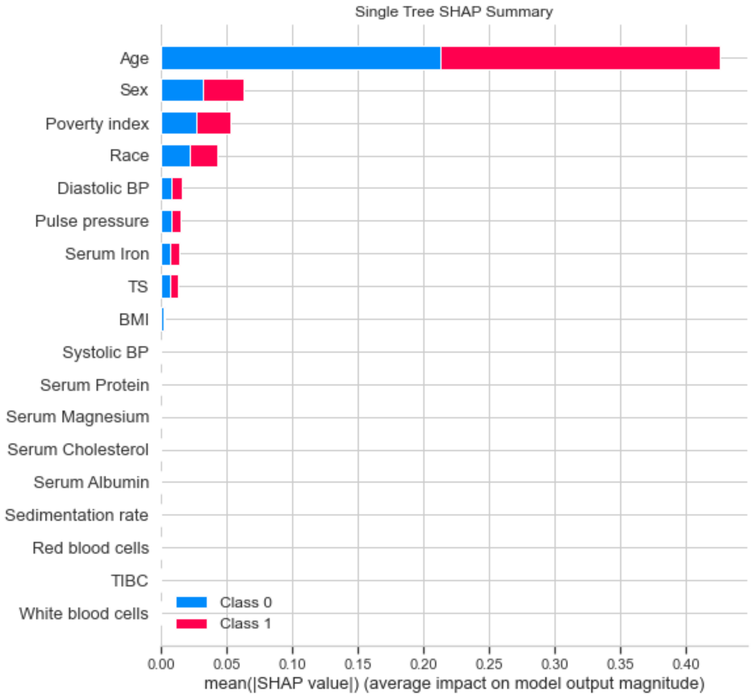

Figure 4: Single Tree SHAP Values Summary

The mean SHAP value rankings indicate that *Age, Sex, Poverty Index, Race, Diastolic BP, Pulse Pressure, Serum Iron, and TS* are all significant features for the bagging model (in order of importance). Moreover, we see from the red/blue breakdown that these features are approximately equally important for predicting survival, as well mortality according to our response variable criteria.

5.3.2 Random Forest

We fit a random forest model including 500 decision trees. We found the optimal hyperparameters by grid searching on the minimum samples to split and the max depth, and then performing 5-fold cross-validation. With the parameters we obtained (max depth 15, min samples split 15), the resulting random forest model had a test score of 0.73.

To gain insight on the significance of the predictors, we considered three different metrics for importances for the fitted random forest model: the built-in feature importance (Gini), permutation importance, and the SHAP values below:

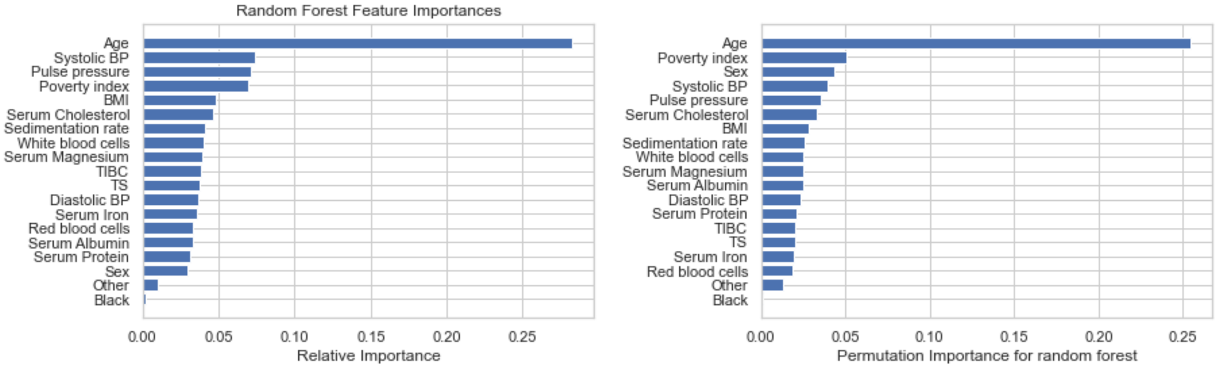

Figure 5: Random Forest Feature and Permutation Importances Plots

What we see is that in all three cases, age was deemed as by far the most important predictor. We also see that all three plots suggest similar importance rankings for the predictors, roughly agreeing on their relative importances with the predictors sex, systolic BP, poverty index, and pulse pressure being close to the top 4 in most cases, suggesting that these predictors are the primary contributors in predicting mortality.

5.3.3 Bagging

For our bagging model, we performed a grid search to find the best hyperparameters (particularly best tree depth from 1-10 and number of estimators from $\{1, 2, 3, 4, 5, 10, 20, 50, 100\}$). The best model produced a ROC AUC score of 0.725.

To glean some insight into what this non-linear and non-parametric bagging model was actually doing, we analyzed the importance of each feature. Recall that the importance of a feature for a decision tree model is computed as the normalized total reduction of the criterion (in this case, ROC AUC) brought by that feature; it is also known as the Gini importance. In this case, since we have a bagging model that aggregates individual tree predictions, we elected to take the mean across each individual tree's feature importance, while also including the standard deviations. We also calculated the permutation feature importance of the bagging model to corroborate our initial findings. Below, we plot the permutation and feature importances of this model. Note that *Age, Poverty Index, Sex, and Systolic BP* are the 4 most informative predictors:

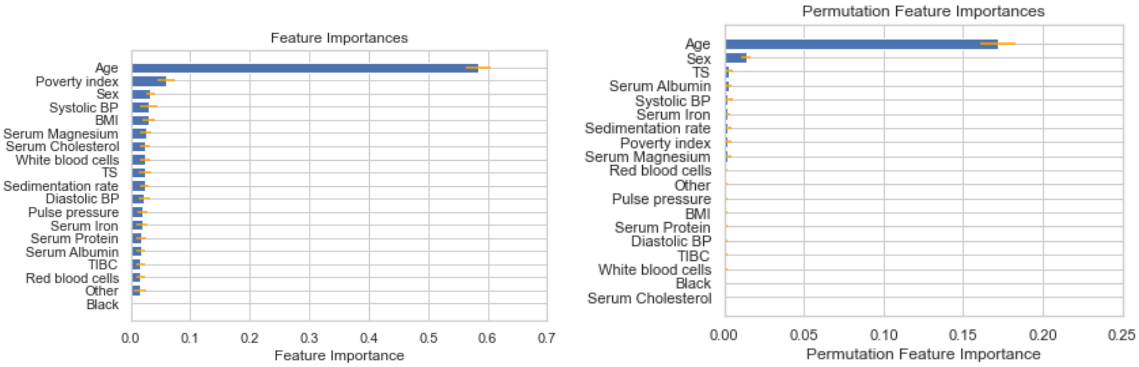

Figure 6: Feature and Permutation Importances for Bagging Model

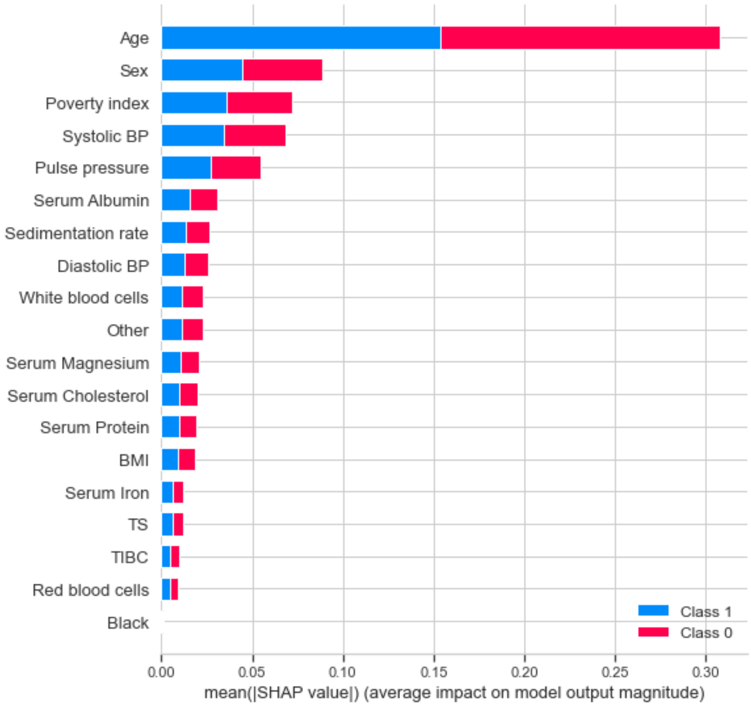

Figure 7: Random Forest SHAP Summary Plot

5.3.4 Boosting

Finally, for our boosting model, we used both AdaBoost and XGBoost for our classification problem of predicting individual survival for at least 18 years from the date of examination. In order to find the best parameters, we plotted AdaBoost accuracies over increased number of iterations along with different max depths of the base learner tree, finding that 120 iterations with a max depth of 3 was the best parameter set via inspection. In addition, we kept our learning rate at 0.05. Our final AdaBoost model yielded an AUC of approximately 0.746, and our final XGBoost model yielded an AUC of approximately 0.745. These AUC scores were the highest of all our models, establishing our boosting model as our best model. Since XGBoost lended itself more easily to visualization than AdaBoost and had very similar AUC, we will consider feature importances for XGBoosting from here.

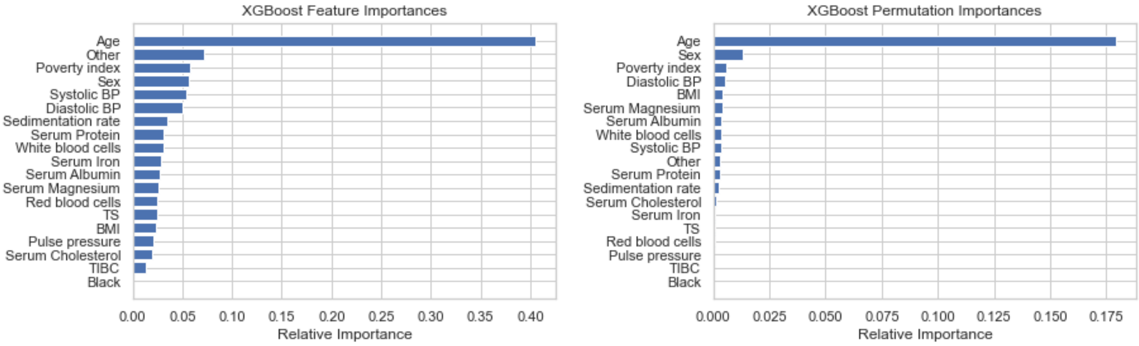

Figure 8: Boosting Feature and Permutation Importances

The above Permutation and Feature Importances reflect slightly differing levels of importance between certain predictors with feature importances reflecting the most important features as *Age, Other, Poverty Index, Sex, Systolic BP, and Diastolic BP* while permutation importances shows *Age, Sex, Poverty Index, Diastolic BP, and BMI* as the most important features. To better interpret this non-parametric model, we found SHAP values along with SHAP interaction values for the top 4 predictors ranked by SHAP values, which as seen by the graph below were *Age, Sex, Poverty index, and Systolic BP.*

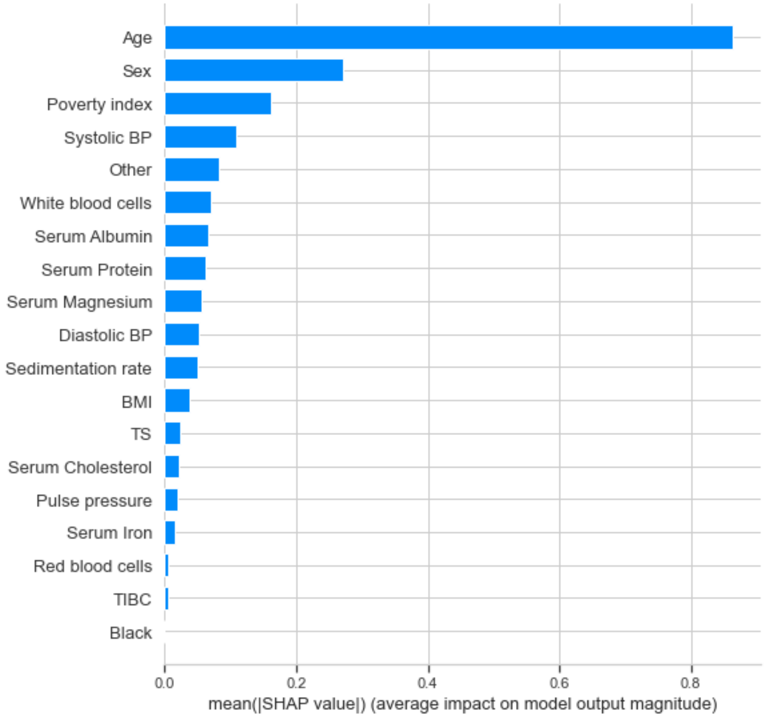

Figure 9: Boosting SHAP “Importances" plot

Since boosting was our best model, we also tested ***feature subsets*** to find a set of features that produced a similar accuracy to the full model. In addition to the full feature set, we also tested our boosting model with different subsets of our features, specifically a demographic only, serum only, iron-related only, and top 4 SHAP features subset. With this, we found that the top 4 SHAP features subset had the highest subset AUC of 0.738 and the demographic subset was close with a score of 0.730 while the other subsets were much lower. Thus, we can say that a subset of Age, Sex, Poverty Index, and Systolic BP features is the smallest feature subset that roughly maintains model performance on a full feature set. The demographic-only subset is also a useful subset if a medical professional is constrained to only having limited records about a patient.

Finally, having seen the top 4 predictors via SHAP for the best boosting model that we had, we created interaction plots between these 4 predictors to see how our model predicts probabilities of survival for at least 18 years from first examination with varying levels of each of these features. This gave the following 6 plots:

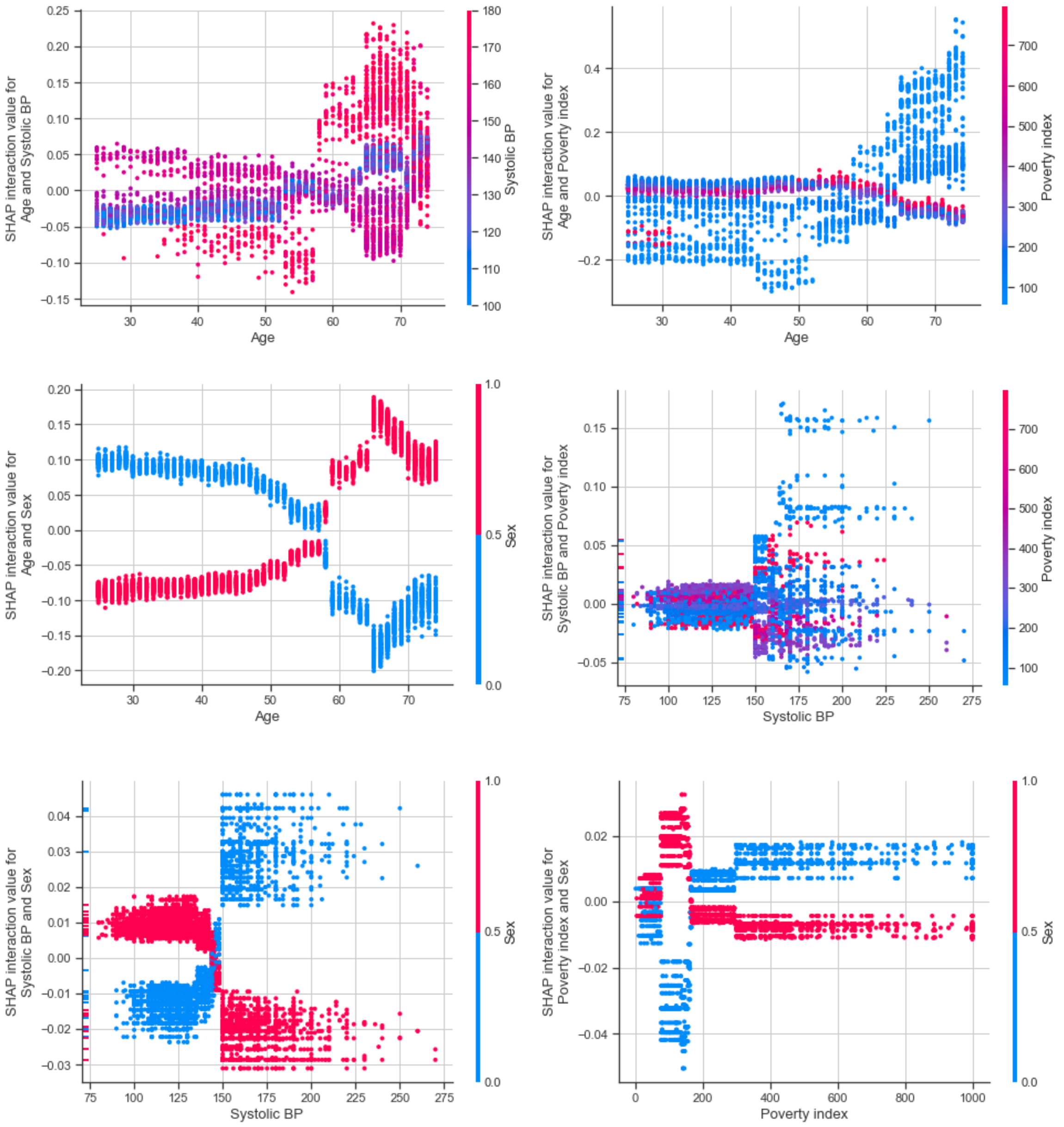

Figure 10: SHAP Interaction Plots for Age, Sex, Systolic BP, and Poverty Index

The top-left plot of Age and Systolic BP reflects that at high ages, some observations with higher Systolic BP had higher predicted probability of survival for at least 18 years. This can be explained via medical reports[1], which conclude that "older frail adults might benefit from slightly higher blood pressure." The top-right plot of Age and Poverty Index suggests that at higher ages, a lower poverty index relates to increased likelihood of survival. This is likely because lower indexed people die at a young age due to poverty-related factors (i.e. low access to healthcare), so those who *do* survive to an older age may bare some external factor

[1] https://www.health.harvard.edu/blog/blood-pressure-goals-may-need-to-change-with-age-201207205034

(such as an abnormally strong immune system) that differentiates them. The middle-left plot involving Age and Sex shows that the sex-based death risk gap varies by age and peaks at age 65, with males having a lower chance of surviving. The middle-right plot of Systolic BP and Poverty Index reflects that some observations at low poverty levels that have higher systolic BP are predicted to have higher survival. Otherwise, the observations are relatively similar in predicted survival across varying systolic BP and Poverty Index. The bottom-left plot of sex and systolic BP reflects that higher BP males and lower BP females have higher predicted survival than their counterparts. The bottom-right plot of Poverty index and Sex reflects that higher poverty level males and lower poverty level females have higher predicted survival than their counterparts with a significant sex-based gap at low poverty index levels. The last three interactions seem to be counter-intuitive, possibly reflecting that there exists some multicollinearity in our data causing the counterintuitive interaction plots and trends.

6 Conclusion

Having trained and tested models from a baseline kNN to complex Random Forest and Boosting algorithms, we found slight increases in AUC score as the model became more complex with Boosting having the highest AUC score, and thus, boosting was the model that we used for our visualization with SHAP and for subsetting features. Overall, however, our boosting model only had marginally higher AUC scores in comparison to our most simple models with all of our models ranging between 0.68 to 0.75 in AUC score. Thus, we chose to focus more on interpretability of our best model to help find the best features and best small subset of features to predict survival of at least 18 years following first examination by the study.

Across all models, we see that the most important individual features when predicting an individual's survival include Age, Sex, Poverty Index, and Systolic BP. From basic models like k-NN to Random Forest and Boosting, we see with increasing ROC AUC scores that these predictors are consistently salient when predicting an individual's likeliness of surviving at least 18 years. We also found the the best small subset of the features was in fact the boosting model trained on these 4 best features. In addition, when looking at plots of SHAP interaction plots, we found many interesting and unexpected results concerning how these four features relate to each other and the response variable, such as seeing that sex and systolic BP have interactions such that higher BP males and lower BP females have higher predicted survival than their counterparts. Thus, based on our modeling and interpretation of our models, we recommend the features of Age, Sex, Poverty Index, and Systolic BP as well as their interactions with each other be considered highly when trying to predict mortality and specifically when predicting survival of at least 18 years since first examination.

We have many reservations concerning the conclusions drawn from our model. A major reservation is that our data had questionable initial trends in both the response and various features that could indicate some bias or simple imprecision in the study. Concerning the response, it was highly skewed with the vast majority of values originally around -20, which reflected people alive at the end of the study or rather, people who simply were not recorded as dead. Thus, in dealing with this imbalance in the response, when creating our categorical response using a threshold of 18, we tried to balance the response but could not manage to make them very equalized. In addition, various unnatural trends were exhibited within features such as red blood cell count that had many of the same value. Moreover, when inspecting the race feature, we see that there is a very small proportion of black people, possibly representing some imbalance or inherent bias within the data collected. Furthermore, we are limited to the features presented in the data set, even though there are possibly other features that are more useful for predicting mortality and could also confound the existing relationships within our model. Finally, this data set is over 30 years old and is possibly not reflective of current trends of causes of mortality. Overall, despite our final findings, we recommend proceeding with caution with the predictions gleaned from our model.

7 Impact Statement

The ability to predict the likelihood of mortality within a definite time span (in this case, 18 years) for an individual would be incredibly beneficial. Indeed, it would allow health professionals and medical experts to prioritize and preemptively aid individuals deemed to be most at risk of dying. Without a doubt, effective early care and treatment for high-risk individuals has a non-trivial possibility of saving and/or prolonging patients' lives.

Furthermore, we believe that it is critical to focus on analyzing, interpreting, and understanding our underlying models' behaviors. It is one task to be able to obtain good accuracy with a black-box algorithm; it is entirely another to understand *how* it works, and how it can be applied in a practical setting, with real-world medical staff dealing with real-world patients. We do not wish to ask a doctor to blindly accept the outputs of a black-box model; we want the doctor to have the ability to observe the most salient features of a model and use their own expertise to gauge the validity of that prediction. In a similar vein, obtaining interpretable models that allow professionals to discern the most critical features in predicting mortality risk is a boon to medical efficiency: instead of gauging risk from a whole host of complicated factors, our models are able to, in a sense, “preprocess" and reduce the work and time that a doctor would otherwise need to commit to.

Ultimately, we hope that our interpretable models are a step towards increased medical efficiency, better targeted preemptive care, and longer lives!

8 Appendix

. . . .

8.1 SHAP Theory

At a high level, a model's SHAP values are the Shapley values of a conditional expectation function of the original model. Under this framework, the importance of a feature is captured by comparing what a model predicts with and without the given feature.

More formally, SHAP values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value $E[f(z)]$ that would be predicted if we did not know any features to the current output $f(x)$. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the ϕ_i values across all possible orderings.

Concretely, the SHAP value for a model f, prediction x, and feature i is given by:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \backslash i)]$$

Here, z' is as subset of “present" features out of the set of entire features x' and M is size of the feature set. In general, computing SHAP values is a NP-hard problem; fortunately we can leverage the power of the **shap** package from Scott Lundberg to approximate these values. We omit the details of these approximation procedures in this discussion.

8.2 Full-Scale EDA Plots

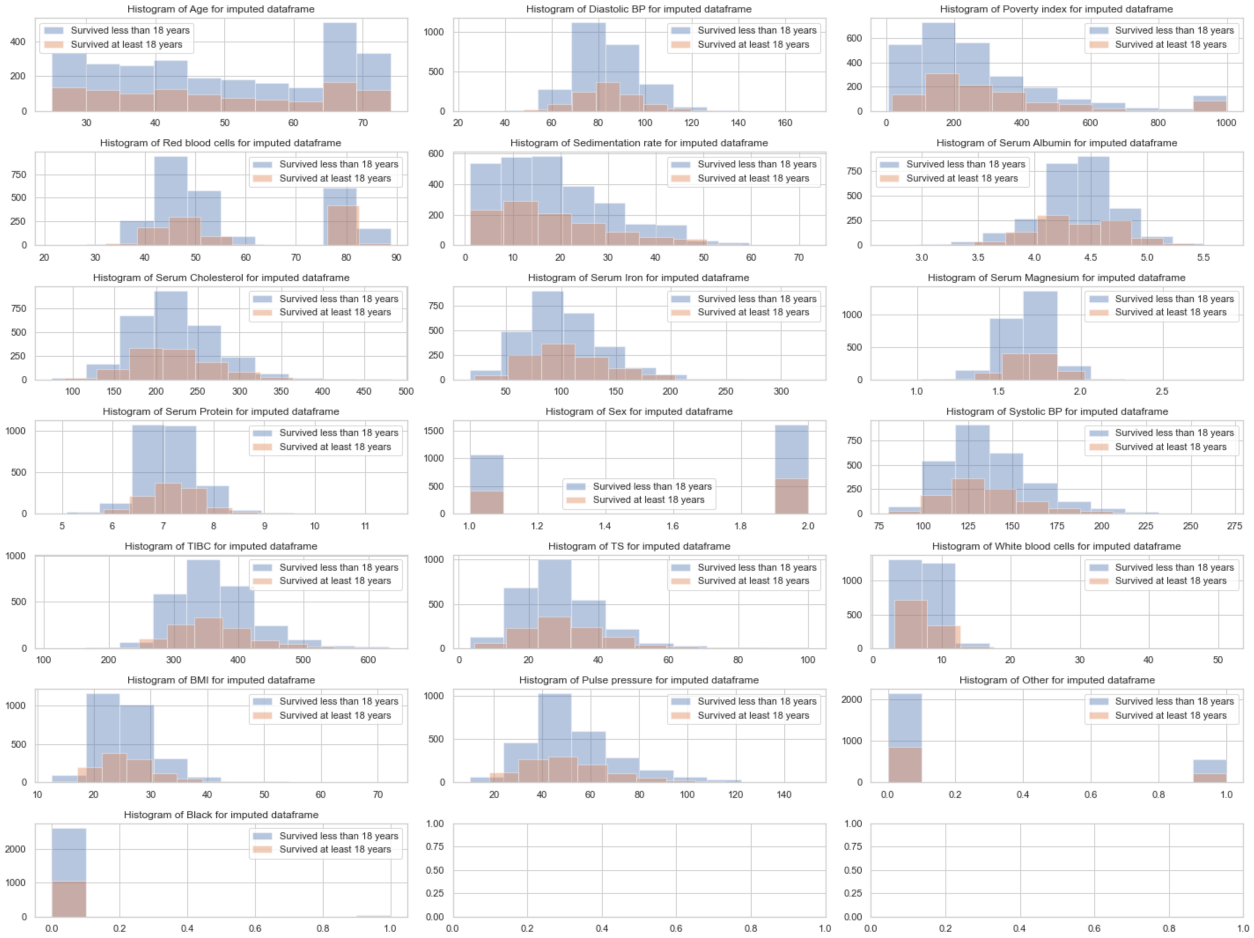

Figure 11: Histograms for Each Predictor

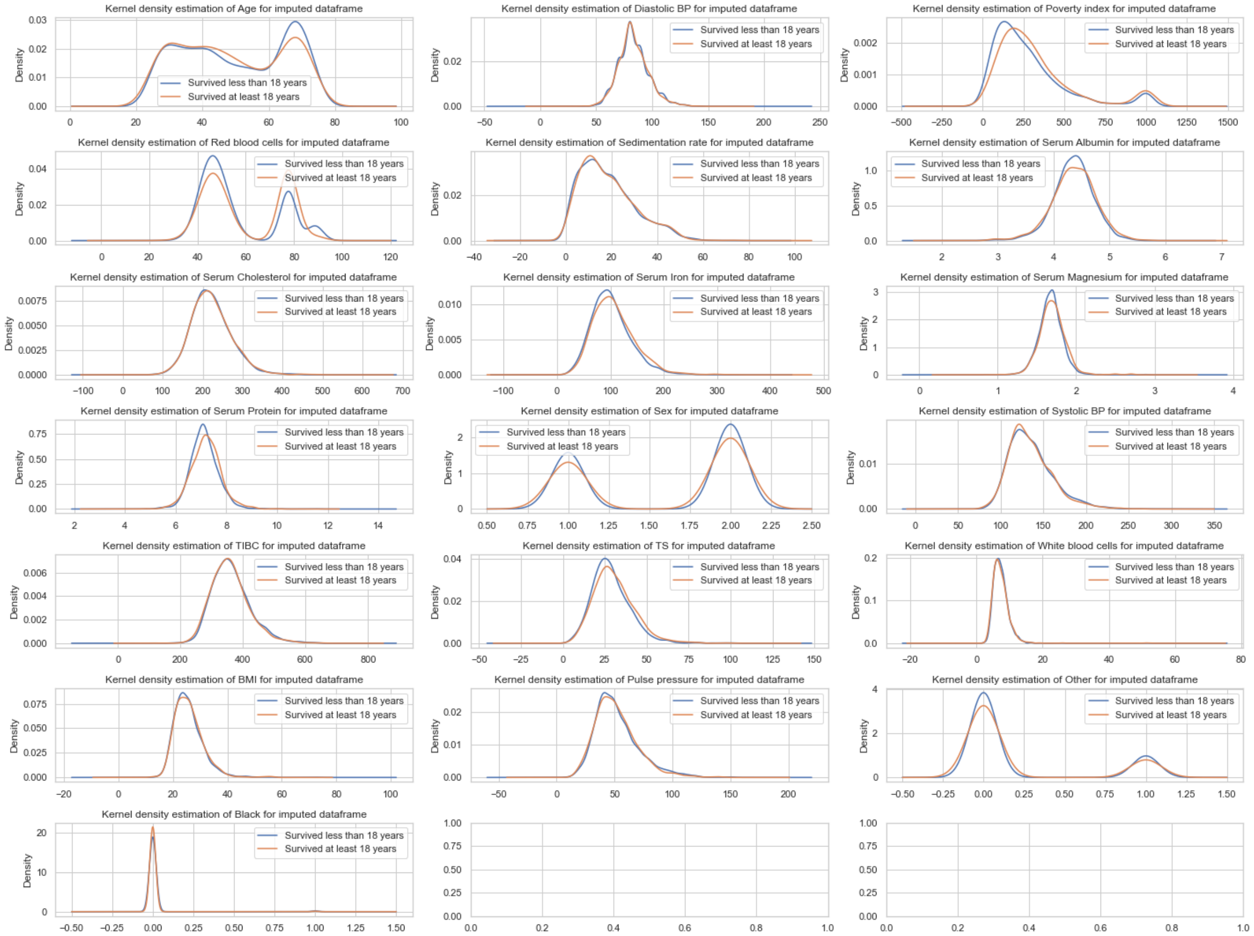

Figure 12: Kernel Density Plots for each Predictor

8.3 K-Nearest Neighbors Plots

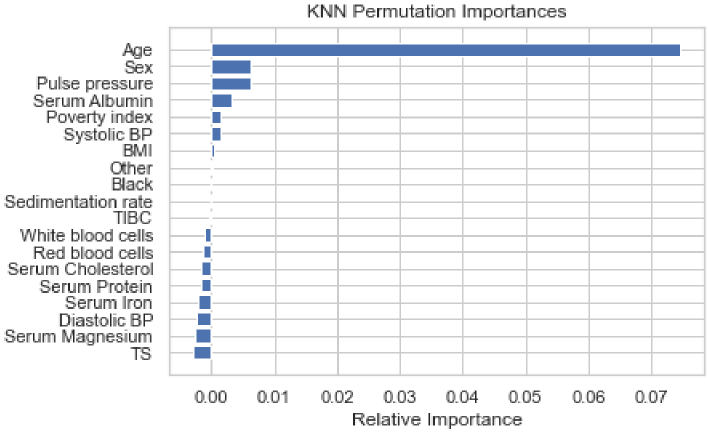

Figure 13: Permutation Importance for k-NN