Machine Learning Applications to Predict Mortality

Authors: Aloysius Lim, Molly Liu, Austin Nguyen

1 Summary
Mortality is a key concern at a population-level as well as at an individual-level. In this paper, we provide three approaches to understanding mortality: (1) model that predicts as a binary outcome whether a person will survive for at least 5 years, (2) a model that predicts whether a person will survive for at least 20 years and (3) a cox proportional hazard model to predict general survival period. We find that the best model for the first two approaches is XGBoost given the highest AUC score of 0.80, a common metric used in machine learning applications. We hope our findings provide health policymakers as well as physicians with additional information to guide their policy decisions as well as healthcare guidance**.**

2 Problem Statement
Longevity and mortality are key metrics that indicate a nation's overall well-being. Longevity and mortality have the potential to look very different across countries, even among countries with similar levels of industrialization and development. In the United States, cardiovascular disease is the leading cause of death for men, women, and people of most racial and ethnic groups, according to the Center for Disease Control & Prevention (CDC). In fact, one person dies every 36 seconds in the United States from cardiovascular disease. About 659,000 people in the United States die from heart disease each year—that's 1 in every 4 deaths. Heart disease costs the United States about $363 billion each year from 2016 to 2017. This includes the cost of healthcare services, medicines, and lost productivity due to death.

Given the prominence of cardiovascular disease in the United States, we are interested in understanding mortality in the United States with data from the National Health and Nutrition Examination Survey (NHANES I). Mortality data will prove useful for medical professionals in guiding the care and treatment plans of their patients. With this frame in mind, our project aims to predict a person's risk of dying, survival period based on a set of clinical data and biochemical measurements captured in NHANES I.

3 Objectives
We aim to answer the following questions in this paper:

- How to predict a person's risk of death and survival period based on a set of clinical data and biochemical measurements?
- Which variables are more important in terms of mortality prediction? Demographic variables (income, race, gender) or clinical results?

To answer the following questions, we develop interpretable models that can be used to provide guidance to medical professionals to direct patient prognosis.

4 Exploratory data analysis
The National Health and Nutrition Examination Survey (NHANES I) was conducted on a nationwide sample of 32,000 people ranging between 1 and 74 years of age in the United States. NHANESI intentionally selected a sample such that certain populations believed to be at high risk of malnutrition were oversampled. These groups included low-income individuals, preschool children, women of childbearing age, and the elderly. Among the 32,000 individuals sampled, 31,973 were interviewed and 23,808 were examined. A subset of 3,854 people 25-74 years of age received a more detailed health exam with an additional 3059 adults later examined in 1975 to augment the data, producing a total of 6,913 people who received a detailed medical examination.

The dataset we are exploring for this project is from the NHANES I Epidemiologic Follow-up Study and is available on the National Center for Health Statistics website. The dataset contains 14,407 observations and 42 columns, including the response variable, y. Among the variables in our dataset, we have divided our predictors into four broad categories: 1) demographic variables, 2) blood test metrics, 3) urine test metrics, and 4) body metrics.

Data is right-censored with not having observed the deaths of the participants by the end of the survey period. The y variable contains both positive and negative values. Positive y values represent people who have passed away at the 1992 follow-up and y is the number of years they survived since the initial examination. Negative y values represent mostly people who have survived till the end of the survey period or till the point at which they dropped out of the survey. For this group of people, we would not have information about their death, but only about their survival period.

4.1 Demographic variables

Most demographic variables demonstrate similar distribution patterns for the two groups, except for age which has the exact opposite distribution for people who are dead and alive. The contrasting distribution makes sense in this case because age for the dead group is skewed to the left which implies that people in this group are generally older whereas the distribution for the alive group is skewed to the right indicating that people in this group are generally younger who are less likely to pass

away.

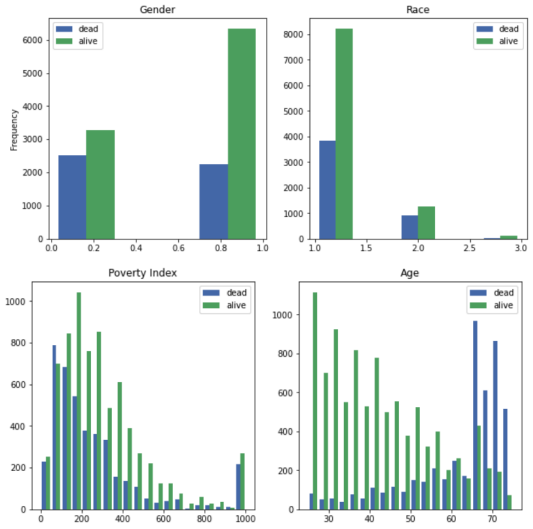


Figure 1: Gender, Race, Poverty Index and Age against survival indicator.

For the blood and urine test metrics, we can see that cholesterol level, uric acid, blood pressure, and pulse pressure are generally higher for the group who have died than for the group that we believe is alive. We will further explore these variables in our model.

4.2 Domain knowledge

Given that the leading causes of mortality in the United States are related to heart disease, we wanted to see whether biomarkers related to heart disease showed a different distribution between the group of individuals who had died and the group of individuals who were still alive.

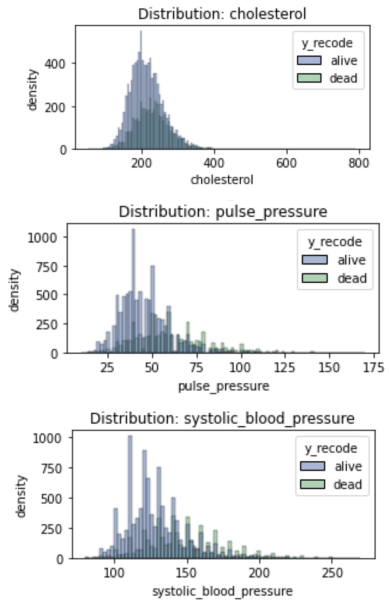


Figure 2: Cholesterol, Pulse Pressure, Systolic Blood Pressure against survival indicator

Across predictors such as cholesterol, pulse pressure, and systolic blood pressure, we do see a right-skewed distribution among the group of individuals who have died compared to the group of individuals who were considered alive. This demonstrates that the NHANES I data aligns with our domain knowledge of mortality in the United States.

4.3 Correlation between response and predictors

Based on initial data exploration and domain knowledge, variables such as age, cholesterol, pulse pressure, and systolic blood pressure are all drivers of mortality. We explored the Pearson correlation between variables, assuming the relationship to be linear in initial explorations. Our correlation plot demonstrates that these variables are highly correlated, posing challenges in the interpretation of each variable as the driver of mortality.

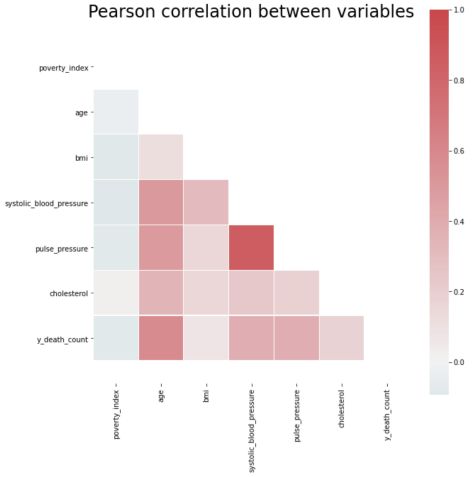


Figure 3: Gender, Race, Poverty Index and Age against survival indicator.

4.4 Data Missingness

AS NHANES is based on voluntary participation, missing data arises when there is the absence of answers to specific questions in the interview. Certain variables have a huge proportion of missing data, evident in the bar chart below showing missing data counts by predictors.

While socioeconomic, and medical history information is collected through household interviews, data related to biochemical measurements are collected from physical examination in mobile examination centers. Hence, there is a strong nullity correlation among data related to biochemical measurements, evident in its nullity correlation heatmap. Data is not missing completely at random. We are expecting data to be missing at random where missing values can be predicted by observable data. For example, missing values in height can be predicted by observable gender and weight.

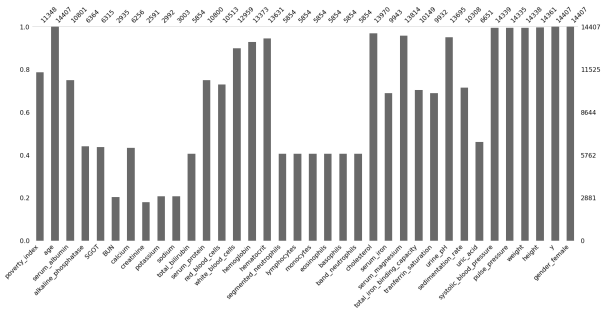


Figure 4: Missingness across all the predictors.

5 Data Processing

5.1 Response

Given that our dataset is a snapshot of an individual's health taken, we choose to model our response as indicator variables for minimum survival duration. In other words, did a respondent survive for at least n years after the NHANES initial examination (i.e. 1 = yes, 0 = no)? In contrast to modeling survival duration outright, modeling this type of binary response for different survival horizons comes with the advantage of being able to return the predicted label as well as the predicted probabilities. The predicted probabilities can then be used to not only estimate patient risk but also rank patients based on risk to allow medical professionals -- who often have little time and bandwidth -- to prioritize their time to focus on providing guidance and interventions to patients with the highest risk of mortality.

We choose to look at survival durations of 5, 10, 15, and 20 years because we are interested in short-term, medium-term, and long-term horizons for mortality. This allows us to separate out individuals likely to survive for at least 5 years from individuals likely to survive for at least 20 years, which are likely separate populations of respondents with different characteristics. For the purposes of downstream work such as missing data imputation, we choose to use survival for at least 5 years as our primary response variable, but we explore all four minimum survival indicators in model selection.

5.2 Missing Data Imputation

Missing data was imputed through an iterative imputation approach by estimating each feature from the others using Bayesian ridge regression. Our strategy is to impute features with the least missing values first. While such an approach could introduce bias into our model, we will attempt to limit such bias by taking into account the proportion of missing values in each feature during our feature selection process. Variables with over 50% missingness are dropped. The 15 dropped variables are all biochemical measurements.

5.3 Feature Engineering and Selection

Given the relatively large number and different types of predictors in our dataset, we performed feature engineering to recode the 6 categorical data through one-hot encoding. We choose to engineer new features that we believe are related to mortality. In the United States, obesity and cardiovascular disease are prevalent and are correlated with mortality. For example, body mass index (BMI) is correlated with all-cause mortality (Baskaran, et al, 2018). We feature engineer BMI from the provided height and weight data following the formula provided by the Division of Nutrition, Physical Activity, and Obesity of the Center for Disease Control & Prevention (CDC).

Given our domain knowledge that cardiovascular disease is the highest cause of death in the United States, we also create new features that capture whether certain cardiovascular biomarkers are within the normal range. These ultimately produce several indicator predictors for BMI (underweight, overweight, obese), systolic blood pressure and diastolic blood pressure (normal, prehypertension, and hypertension), and serum albumin (low, normal, high). In total, our feature engineering resulted in 25 new predictors.

From there, we performed feature selection using 5 different methods: 1) SelectKBest univariate feature selection with chi-square test to weed out the features that are the most likely to be independent of class and therefore irrelevant for classification, 2) L-1 based feature selection so that the higher the alpha parameter, the fewer features selected, 3) Permutation importance, a procedure that breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature, 4) Recursive feature elimination with cross-validation step size 5 to select the number of features, and 5) Tree-based feature selection to compute impurity-based feature importances, which in turn can be used to discard irrelevant features. We combined the feature selection results from all 5 models. Age is selected by all 5 models which are in accordance with the assumption and common sense that it is the most relevant variable in predicting mortality. Features such as poverty index, blood pressure, pulse pressure, urine, gender, and BMI appear in 3 or more feature selection models. The feature selection models dropped some of the highly correlated variables. For instance, the one-hot encoded race variables are not selected because of their high correlation with the poverty index. In the end, we selected 30 predictors that appear in the feature selection model more than once to include in our model.

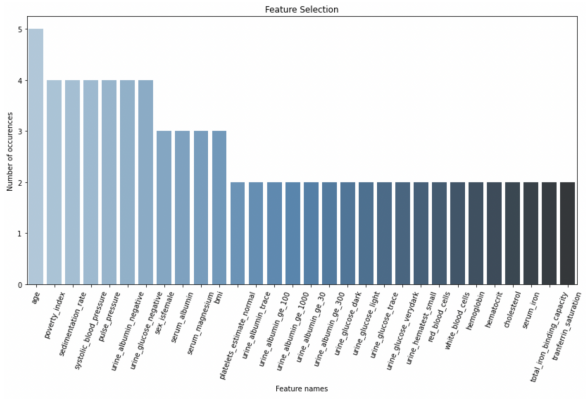


Figure 5: Frequency of predictor occurrence across feature selection approaches.

6 Modelling Approach

6.1 Model exploration

We explored four different models to predict the minimum survival indicators: RandomForest (RF), Gradient Boosted Trees (GBM), XGBoost (XGB), and Logistic Regression (LR). We mainly focus on tree-based models due to their ability to handle non-linear patterns in our data and we treat Logistic Regression with LASSO regularization as a baseline.

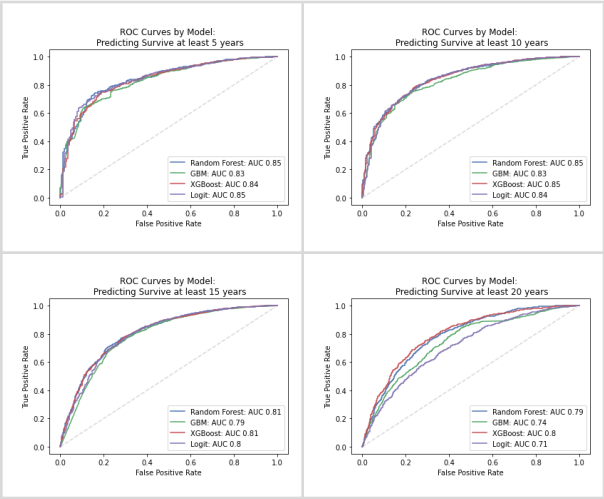


Figure 6: ROC Curve Comparison of RandomForest (RF), Gradient Boosted Trees (GBM), XGBoost (XGB), and Logistic Regression (LR) for minimum survival indicators as our response for 5, 10, 15, and 20 years.

For predicting survival of at least 5 years, performance is largely comparable across RandomForest, GBM, XGBoost, and Logistic with regularization, all yielding ROC AUC values of ~0.84. However, as we lengthen the minimum survival indicators from at least 5 years to 10, 15, and 20 years, we see that XGBoost and Random Forest perform the best. XGBoost showed an ROC AUC score of 0.85, 0.81, and 0.90 across the minimum survival time indicators of 10, 15, and 20 years, respectively. RandomForest showed an ROC AUC score of 0.85, 0.81, 0.79. GBM and Logistic with regularization did not perform as well. Ultimately, XGBoost performs the best on AUC (0.80) for the test set. Notably, the AUC of GBM and Logistic Regression worsen as we extend the time horizon of the survival indicator whereas the AUC of RandomForest and XGBoost perform remain fairly strong.

We also examined Precision-Recall (PR) curves and PR-AUC to understand model performance acknowledging class imbalance. Unsurprisingly, all the models are tightly clustered together when modeling minimum survival of at least 5 years. This is because 93% of the dataset lived for at least 5 years, so the positive class is the majority. However, when modeling a minimum survival indicator of 20 years (in which case ~23% is in a positive class), the differences across the four different models are a lot clearer: RandomForest performing the best with a PR-AUC of 0.88.

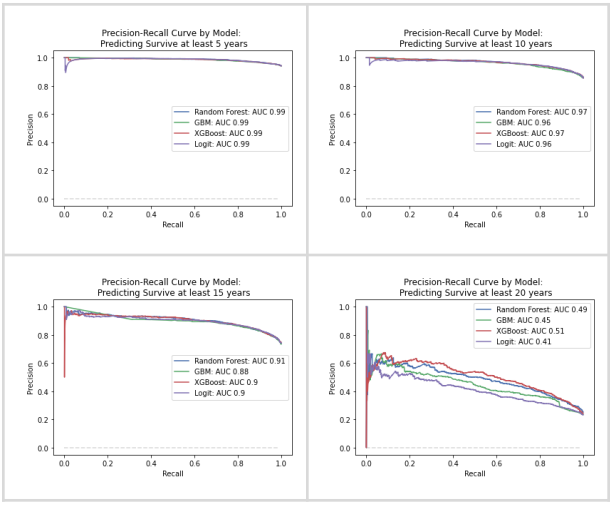


Figure 7: Precision-Recall Curve Comparison of RandomForest (RF), Gradient Boosted Trees (GBM), XGBoost (XGB), and Logistic Regression (LR) for minimum survival indicators as our response for 5, 10, 15, and 20 years.

For the model that predicts a minimum survival indicator of 5 years, performance is comparable across all models. However, as we extend the minimum survival indicator to look at 10, 15, and 20 years, we do see differentiation in model performance. We see the largest differences in models when predicting survival indicators of at least 20 years, with XGBoost performing the best with a PR-AUC of 0.51, followed by RandomForest (0.49), GBM (0.45), and Logistic Regression (0.41). Thus, we selected XGBoost as our final model because it performs best on ROC-AUC and PR-AUC, particularly for minimum survival indicators of 20 years. This illustrates that XGBoost is robust when trying to predict survival indicators across multiple time horizons.

6.2 Model Interpretation for Minimum Survival Indicator of 5 years

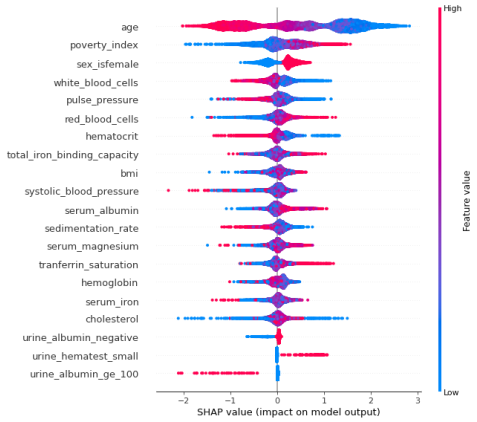


Figure 8: SHAP Force plot that shows features that have a positive impact on the model (pink), negative impact on model (blue), and no impact on the model (purple).

SHAP Force plot highlights that age, poverty index, sex, white blood cells, pulse pressure, red blood cells, hematocrit, and total iron-binding capacity are the top features for our model that predicts whether an individual will survive for at least 5 years.

Acknowledging the correlation between age and many of our predictors, we also examined SHAP dependence plots to understand the contribution of interacting terms to the overall SHAP values. We compared age against systolic blood pressure, pulse pressure, and sedimentation rate.

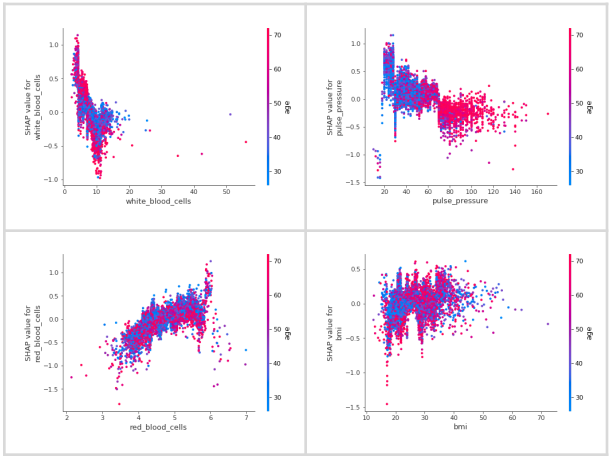


Figure 9: SHAP dependence plots show thresholds that are inflection points in SHAP value across the value of the predictors as well as the relationship to age.

For example, pulse pressure values less than 65 demonstrate SHAP values close to 0, but after pulse pressure exceeds 65, we see a noticeable drop in SHAP values into the negative range. This aligns with our domain knowledge that as blood pressure increases, we see more patients exhibit signs of pre-hypertension and hypertension which are symptoms of cardiovascular disease. The inflection at the value of 65 for pulse pressure corresponds as well as with age as individuals pulse pressures higher than 65 also skew older in age.

All that said, our model is not a causal model. The effect we observe in these SHAP dependence plots may have to do with confounding (e.g. age and pulse pressure could be correlated to another variable not observed in the dataset).

6.3 Model Interpretation for Minimum Survival Indicator of 20 years

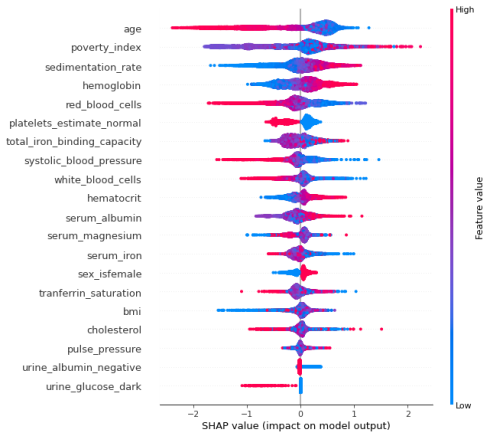


Figure 10: SHAP force plots show that the top 5 features are age, poverty index, sedimentation rate, hemoglobin, red blood cells..

In addition to predicting mortality in the short-term, we are also interested in longer-term outcomes. Like the model that predicts a minimum survival of 5 years, our model that predicts survival of at least 20 years shows that age and poverty index are the most important predictors. However, sedimentation rate and hemoglobin are considered top features and sex drops out as important in this model. Our interpretation is that socioeconomic status via poverty index and age are more indicative of long-term survival outcomes compared to sex.

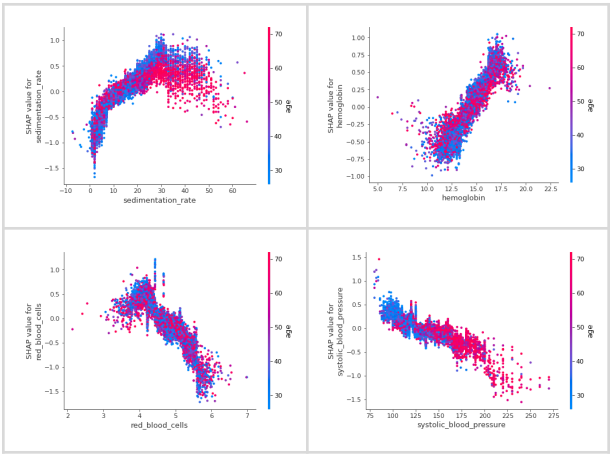


Figure 11: SHAP dependence plot demonstrates inflection points: 125 for systolic blood pressure and 25 for sedimentation rate.

When examining the SHAP dependence plots, we observe inflection points for sedimentation rate (~25), hemoglobin (~15), red blood cell count (5), and systolic blood pressure (125). We can also observe the correlation between age and cardiovascular biomarkers such as systolic blood pressure and pulse pressure, which aligns with our domain knowledge of high cardiovascular risk in the United States.

6.4 Model interpretability: Clinical Context

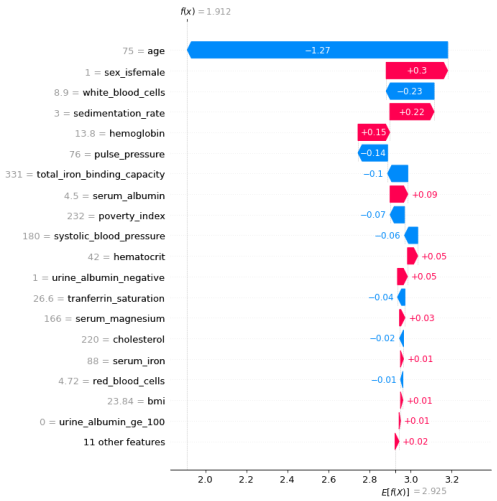


Figure 12: SHAP waterfall plot demonstrates the contribution of each predictor on the model's final prediction. This will look different per patient. By giving medical professionals the contribution of a predictor on an individual's minimum survival outcome, SHAP waterfall plots lends themselves well to clinical applications.

For this specific individual, our model predicts 1.912 log-odds for their minimum survival years after initial examination, which is lower than the base value of 2.925 log odds of surviving for at least 5 years since examination (which represents the average survival time in years across the entire dataset).

This means that this individual's log odds of surviving for a minimum of 5 years is lower than the average survival time across the entire dataset. The individual being 75 years old (relatively old in the dataset) decreased their log odds of surviving the most, followed by being female which had a positive contribution to the individual's log odds of surviving for at least 5 years. This makes sense because older individuals tend to be less healthy than younger individuals and women have longer life expectancies compared to men.

The sum of contributions of all the predictors yields the difference between our model's prediction (1.912) and average prediction (2.925).

6.5 Model performance with fewer features chosen

The SHAP values demonstrate that some predictors have higher importance than others in the model. In order to make our model more interpretable under clinical context, we explore the impact of a decreased number of features chosen on the AUC score. We want to find the smallest feature subset possible while maintaining a relatively good model performance. In order to achieve this, we performed initial feature selection which decreased our predictors by half. We further selected only the top 10 predictors in the previous XGBoost model to obtain a new AUC score of the ROC curves for survival in 5 years and survival in 20 years respectively.

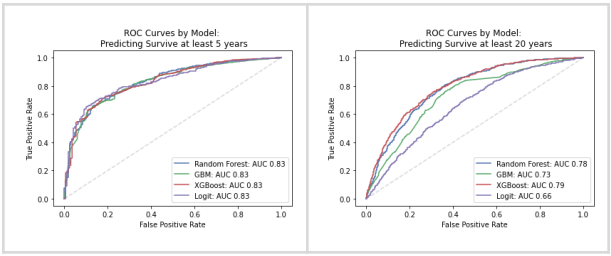


Figure 13: ROC Curve Comparison of RandomForest (RF), Gradient Boosted Trees (GBM), XGBoost (XGB), and Logistic Regression (LR) for minimum survival indicators as our response for 5, and 20 years with top 10 predictors

Comparing the two graphs above with the previous AUC scores with 30 predictors, we find that the AUC scores only decrease slightly for the models with fewer predictors. AUC score decreases from 0.84 to 0.83 for the survive at least 5 years XGBoost model and decreases from 0.8 to 0.79 for the survive at least 20 years XGBoost model. This demonstrates that we can remove some of the features, even up to 60 percent of the predictors, and achieve similar performance. With these findings, when presented with a large number of demographic and biochemical data, doctors can choose a relatively smaller number of most important predictors without compromising the quality of the prediction.

7 Survival analysis

Our analysis so far has mainly been concerned with the survival period. As the data captures both the death event and survival period, we have looked into survival models, specifically the Cox Proportional Hazard model to get the most statistical information out of the data.

7.1 Cox Proportional Hazard Model set-up

The responses of our modeling are both the survival period, and the event of interest is death. Hence, the hazard rate measures the probability of death, given that the participant has survived up to a specific time.

Based on earlier modeling, we have seen that there is a general consistency in the important variables influencing the probability of surviving, which corresponding would influence the probability of death. A smaller subset of the feature subset may be useful in clinical practice, and also

not to unnecessarily capture spurious regression results. We have referenced earlier results in selecting 10 predictors for survival analysis.

Penalty to the size of the coefficient was imposed to control for the stability and correlation of the predictors. After imposing a penalty of 0.1, the p-values of most of the coefficients are below 5%. Breslow is adopted as the baseline estimation.

7.2 Interpretation

The exponential coefficients are the hazard ratios of each predictor, per unit increase in predictors. A higher hazard indicates a higher probability of death relative to the baseline hazard rate.

Using Hazard ratios and due to the nature of the scale, age does not look like the most important variable. An increase in age by 1 year increases the probability of dying by 5% while being female reduces the probability of dying by 34% relative to the baseline.

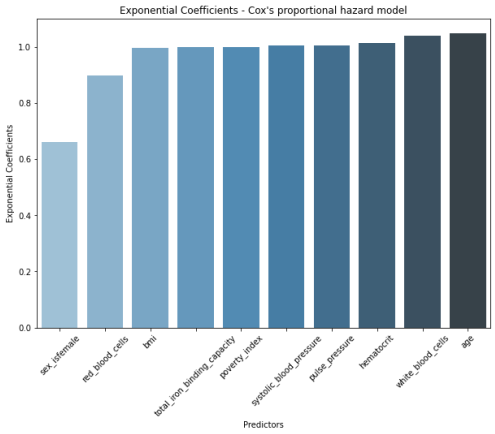
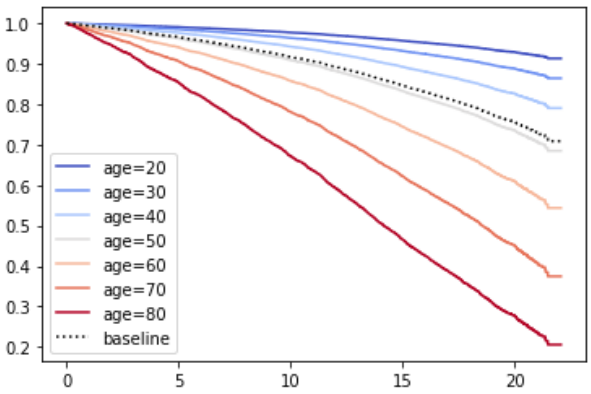


Figure 14 : Exponential coefficients for predictors in Cox's proportional hazard model

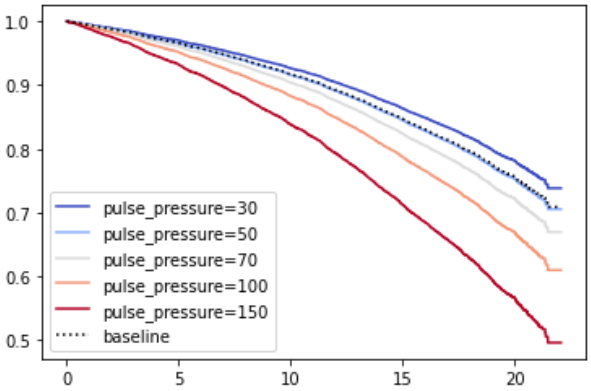
The following graphs show the survival curves as we vary a single predictor while keeping all other predictors constant. The survival curves represent the probability of survival over a time period. We have chosen four results of four predictors to illustrate the interpretability of the model. For ages, the decrease in survival probability decreases in greater proportion for every ten years of increase in age. One noteworthy finding is that BMI, after taking into account other correlated factors, does not have a big impact on survival probability.

A doctor could use the survival graphs of predictors to estimate the survival curve of an individual patient to achieve optimal health outcomes. For example, a doctor could use the survival curve by pulse pressure to advise on the survival probability of a person with a high pulse pressure of 150. In about 20 years, the probability of survival would drop to roughly half of the current level. This can support doctors' recommendations and motivate health-changing behaviors.

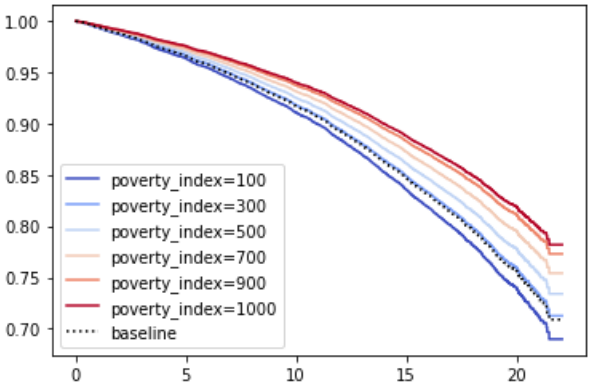
By ages



By pulse pressure



By poverty index



By BMI

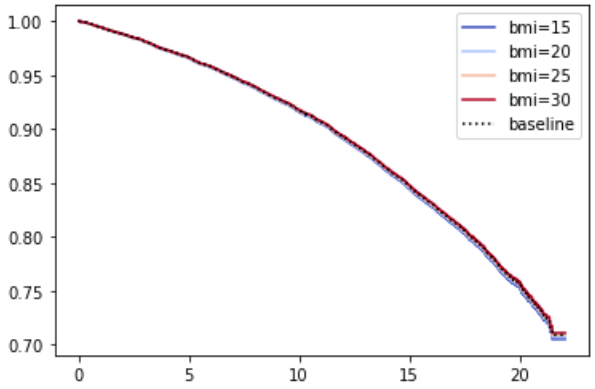


Figure 15: Figure X: Survival curves of varying selected predictor while keeping all other predictors constant

8 Strengths and Limitations

8.1 Consistency of results and interpretability

Our research aims to provide doctors with a straightforward diagnosis of whether a patient would survive in a given period of time and the survival period given a series of demographic and biochemical measures. In order to make our approach more interpretable, we include the most influential predictor variables in our modeling process. Throughout the exploration of feature selection and complex modeling, we have yielded relatively consistent results across various models with similar levels and trends of AUC. Furthermore, variables that have the most significant impact in predicting a patient's mortality have been relatively consistent throughout the research. The top predictor is age which is in accordance with our assumption, but we have also explored many other important factors such as gender, poverty level, blood pressure, pulse pressure, and BMI. Although many of the biochemical measurements are correlated with age, our initial feature selection dropped some of the highly correlated variables and we have created new predictor variables based on a literature review to increase modeling accuracy.

In terms of model interpretability, we want our target audience to gain insights from our research instead of simply providing a model with a high accuracy score. As a result, we explored SHAP to clearly convey how each predictor impacts mortality and gave an example with clinical context to illustrate which demographic or biochemical predictors potentially influence the survival years to what extent.

8.2 Mortality improvement

With the technological advances in medicine, there has been a general trend of mortality improvements that have not been explicitly allowed for in our modeling. Based on the Societies of Actuaries' published report on mortality improvement, there is an overall improvement of approximately 1.1% per year from the survey start period (1971-1974) to 2015. Hence, the conclusion drawn from our modeling would be overstating overall mortality by the same amount and needs to be correspondingly adjusted for.

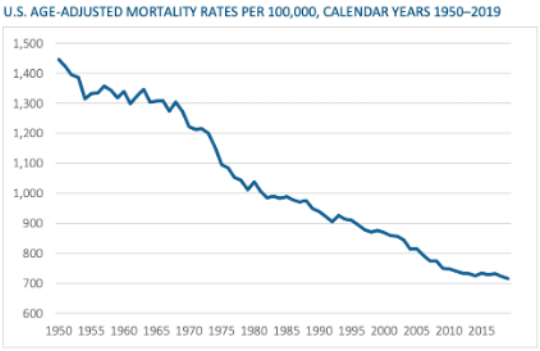


Figure 16: U.S. Adjusted Mortality Rate per 100k people. Source: Society of Actuaries Mortality Improvement Scale Report Figure

Mortality improvement differs by causes of death. For example, the mortality improvement for heart diseases is different from mortality improvement for cancer. Hence, our adjustments to our conclusions regarding heart-disease-related predictors (i.e. systolic blood pressure) should consider the differences in improvements.

8.2 NHANES data – High-risk population

The NHANES I sample was selected so that certain population groups thought to be at high risk of malnutrition were oversampled at pre-set rates. Consequently, conclusions drawn are more applicable for prediction on similar high-risk groups and less applicable on lower-risk groups.

For conclusions to be drawn on a separate group, appropriate sample weights have to be applied within our modeling. This is so that the weighted data would better reflect the characteristics of a particular group where we make predictions on.

Weightings would however not fully correct the bias if data for such groups have not been captured at all. For example, if the data collected did not include healthy individuals for the age group 80-85, changing weightings would not help us draw conclusions for this particular age group.

9 Conclusions

Tying to our objective of finding out which are the most important predictors of mortality, our models generally produced consistent results. Age is unsurprisingly the most important variable with gender, pulse pressure and poverty index, and a few other blood-related variables

showing up as the other important variables. As such, we would recommend that such information should be collected for predictions on survival or mortality.

On our interpretability objective, a one-size fit all approach to modeling is unlikely able to address all questions from users. As results interpretation depends on our modeling approach, any such interpretation should be based on the objective of the users.

For example, an actuary may only be interested in the survival period of 5 years for the design of an insurance product. Interpretation should then be based more on the modeling results of the binary outcome of 5-years survival. For a doctor who may be interested in the survival curve of a patient for the next 20 years, interpretation should then be based on our survival analysis.

We would strongly recommend that any actions decided based on the results of our modeling should only be made with a good understanding of our limitations, and correspondingly adjust for any limitations. In addition, we should be mindful of ethical considerations relating to unfair outcomes from discrimination. We believe that there is scope for future work which could improve the usability of the work as described in the next section.

10 Future work

10.1 Predictors with missingness

For our work, we have focused on predictors with less missingness and limited bias on prediction from missingness. The approach however does not allow us to answer questions relating to these predictors. For example, we cannot answer how the measurement of blood-test-related variables relates to mortality.

While analyzing missingness, we noticed that the missingness of some predictors is strongly correlated i.e. the blood-test-related predictors. These could be due to the collection of data for all such predictors for the same tests.

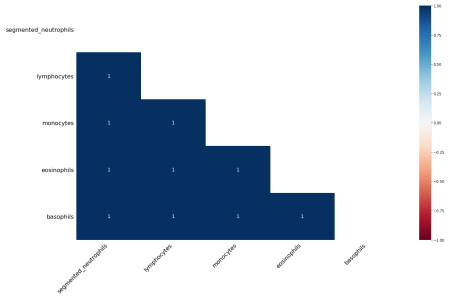


Figure 17: Nullity correlation of selected blood-tests related predictors:

Inferences from the results on such blood tests may be of interest. To draw inferences from these predictors, we can perform detailed analysis on a subset (40-50%) of the data where information from these predictors is jointly available.

10.2 Extracting value of non-age predictors by modeling excess mortality by ages

Age is naturally the most important predictor of mortality, which is also reflected in our results. It may be of interest to remove the effect of aging on mortality. Excess mortality by age can be defined as the difference between the observed mortality and the expected mortality by age. Modeling excess mortality can better provide information about the burden of mortality related to other predictors rather than age.

We can first introduce the expected mortality by considering historical population mortality rates by age over the same period. One example of such population mortality rates can be found on CDC Life Tables. Using these mortality rates, we can then calculate the expected mortality.

(1) Expected mortality = q_x x exposure in life years

where *q_x represents the mortality rate for age x.*

We can then calculate:

(2) Excess mortality = Actual mortality - Expected mortality

The modeling approach will have to correspondingly adjust to the change in the target variable.

10 References

Heart Disease Facts | cdc.gov. (2021, September 27). CDC. Retrieved November 29, 2021, from https://www.cdc.gov/heartdisease/facts.htm

Baskaran, et al. 2018: https://www.thelancet.com/journals/landia/article/PIIS2213-8587(18)30288-2/fulltext

CDC Life tables https://www.census.gov/content/dam/Census/library/publications/2020/demo/p25-1145-supplemental-tables.pdf

Society of Actuaries Mortality Improvement Scale Report https://www.soa.org/globalassets/assets/files/resources/experience-studies/2021/2021-mp-scale-report.pdf