Mortality Prediction and Feature Selection Using NHEFS Data

Xin Zeng[1], Xiaoxuan Liu[2], Benjamin Liu[1], and Yujie Cai[1]

[1]Graduate School of Arts and Science, Harvard University
[2]T.Chan School of Public Health, Harvard University

December 13, 2020

Abstract

Predicting the probability of the occurrence of a binary outcome or event is of key importance in many areas of clinical and health services research. We utilize models in the essence of Logistic Regression and GBDT model to obtain the around 84% accuracy in predicting the mortality rate. Furthermore, by examining the SHAP value of our GBDT model using LightGBM framework, we discover the smallest subset of predictors to achieve a comparable performance. We find out that age, sex, systolic blood pressure, and sedimentation rate is the most robust while keeping high accuracy in predicting the mortality rate.

1 Introduction and Literature Review

Predicting the probability of the occurrence of a binary outcome or event is of key importance in many areas of clinical and health services research. Accurate prediction of the probability of patient outcomes, such as mortality, allows for effective risk stratification of subjects and for the comparison of health care outcomes across different providers.

Factors that affect the mortality
A plethora of medical and statistical literature has investigated the factors affecting the risk of death using the NHANES data set. For example, it is tested by Greenberg (2001) [5], increasing mortality risk is associated with increasing average-adulthood Body Mass Index, down to the lowest BMI category ($< 20 kg/m^2$). Besides, Garg et al. (2002) [4] split NHANES I predicted variables like systolic blood pressure, diastolic blood pressure, serum cholesterol, and serum albumin into group, and show that these variables have different predictive power in cardiovascular risk. What's more, Veeraana et al. (2013) [13] discuss that the abnormal red cell distribution width, level of trasnferrin and serum ferritin, also the erythrocyte sedimentation rate indicate the increase in the chance of getting coronary heart disease, from 1999-2004 NHANES data.

Models used to predict mortality
The logistic model and the decision tree are the most used ones in modeling the mortality rate. Austin et al. (2011) [1] use regression trees, random forests, boosted tress, and logistic regression on the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study. They conclude that bagged regression trees, random forests and boosted regression trees may result in superior prediction accuracy. Other studies [4] [13] with panel data will often time conduct survival analysis, for predicting the mortality.

Metrics used in model evaluations
The common evaluation metrics presented in the medical literature are Hazard Ratio [3][13][4][9]. Another most used ones are the receiver operating characteristics (ROC) curve and the area under the curve (AUC) by Austin et al [1] and Thuluvath, Yoo and Thompson (2003) [11]. In terms of our research, we choose mainly the AUC, together with a more intuitive measure of prediction accuracy to be our metrics, since these two are more generic in data science community.

We use a pre-processed dataset generated from the subset of NHANES I Epidemiological Follow-up Study (NHEFS) data from National Center for Health Statistics to model for the mortality rate. In the light of literature, we start our modeling in our research from a baseline logistic regression and decision tree, gradually adding elements in the regression, and extending to ensemble-based methods such as boosted trees. Choosing prediction accuracy and AUC as our metrics, we compare the performance of models. Moreover, we conduct feature selection and manage to find the subset of features to have the most predictive power on mortality rate, as well as to have a comparable performance with the model with all classifiers.

2 Data

The data comes from a pre-processed dataset generated from the subset of NHANES I Epidemiological Follow-up Study (NHEFS) data from National Center for Health Statistics. There are 9932 observations and 18 predictors in the dataset including medical survey information and demographic attributes. Table 1 gives a description of all predictors in their unprocessed form. It includes the definitions of predictors retrieved from Center of Disease Control and Prevention NHANES official website [8]. We produce summary statistics in the table for each predictor. Notice that the given dataset takes two kinds of variables, laboratory and demographic, for analysis. The feature variables are all numerical, except for *Race* and *Sex* which are categorical.

The outcome of this study is interpreted in 2 parts. The positive values indicate the number of years a person lived after the end of study. The negative values indicate the number of years a person died or the record was lost before the end of study.

2.1 Missing Data

Based on the number of missing data, we believe that they do not pose a severe problem. The independent variables that have missing data are: *Diastolic BP, Sedimentation rate, Systolic BP, White blood cells*, and *Pulse pressure*, among which, *Pulse pressure, Diastolic BP*, and *Systolic BP* miss less than 0.6% of the data, *Sedimentation rate* misses 8.32% of the data, and *White blood cells* misses 10.48% of the data.

We choose to perform multiple imputation to handle the missing data for the following reasons. The data is likely missing not at random. Notably, *Pulse pressure* is calculated from the difference between *Diastolic BP* and *systolic BP*, so the missing is inter-correlated and we use multiple imputing to learn this relationship for imputation. For *White blood cells*, which has the most missing data, the missing is not non-trivial, so multiple imputation can also be more robust in this case.

Table 1: Raw data descriptions

Variable	Description	Mean	St.Dev.
Diastolic BP	Blood pressure, range 25-180, in *mmHg*	83.2819	13.2919
Systolic BP	Blood pressure, range 80-150, in *mmHg*	134.8544	24.9320
Pulse Pressure	Difference of DBP and SBP, in*mmHg*	51.5755	18.2990
Red blood cells	Nucleated red blood cells, count, unit unsure	54.8805	14.6007
White blood cells	Nucleated white blood cells, count,unit unsure	7.4517	2.2920
Sedimentation rate	Measures how fast red blood cells fall to the bottom of a tube, checking for inflammation, in *mm/hr*	16.2682	11.5117
Serum Albumin	Test result, in *g/dL*	4.3651	0.3312
Serum Cholesterol	Amount of cholesterol in the blood, in*mg/dL*	221.2348	49.4944
Serum Magnesium	Test result, unit unsure	1.6813	0.1459
Serum Protein	Test result, in *g/dL*	7.1040	0.5089
Serum Iron	Test result, unit unsure	101.1172	37.2116
TS	The value of serum iron divided by transferrin, in *%*	28.5559	11.2326
TIBC	measures the blood's capacity to bind iron with transferrin, in *umg/dL*	362.6045	58.9936
Age	Age of participants, from 25-74	49.4490	15.8784
Poverty index	Ratio of monthly family income to the household poverty guidelines specific to family size	287.2130	223.3104
Race	1: White, 2: Other, 3: Black	N/A	N/A
Sex	1: Male, 2: Female	N/A	N/A
BMI	Body mass index, in kg/m^2	25.6897	5.1841

Source: Epidemiological Followup Study (NHEFS), 1892-1894, retrieved in 2020

2.2 Data processing

Independent Variables

We recognize that many medical papers (as mentioned in the introductory section) group some laboratory variables into normal and abnormal categories, *Age*, and BMI into different levels. We choose to perform the grouping on *Age*, *Diastolic BP*, *Systolic BP*, *Pulse pressure*, *Sedimentation rate*, *Serum Cholesterol* since these grouping are mentioned in the literature and worth trying in comparing the model performance. We group

based on the criteria shown in Table 2.

While we transform a large portion of the predictors into categorical variables, we still retain the numerical raw data. In the exploratory data analysis stage, we will use both the grouped data and raw data to conduct initial explorations and build visualizations. The version of the data that we will be using in our models will be decided in the modeling stage.

Table 2: Selected grouping of independent variables

Model	Parameters
Age	< 45 young, $45 - 60$ Middle Age,> 60 Elder [3]
Diastolic BP	< 90 normal, 90-120 high, > 120 hyper [12]
Systolic BP	< 140 normal, 140-180 high, > 180 hyper [12]
Pulse Pressure	< 40 low, 40-60 normal, > 50 high [2]
Sedimentation rate	< 14 low, 14-29 normal, > 29 high [13]
Serum Cholesterol	< 230 normal, > 239 abnormal [4]

Binary Response Variable

In the raw format, a positive response variable denotes the number of years that a person still lives after the end of this study, while a negative response variable denotes the number of years a person died before the end of this study or the number of lost-to-follow-up years before the end of this study. Due to the information that the response variable encodes, we have converted the response variable into a binary variable (denoted *post-study mortality*) based on its positiveness because we believe that whether someone died before or after the end of study is more meaningful than the numerical values themselves. We let *post-study mortality = 0* for negative response values and *post-study mortality = 1* for positive values.

2.3 Explanatory data analysis

We perform a 80-20 train-test split on the original dataset and conduct the exploratory data analysis (EDA) on the training set. We focus on the relationship among *Race*, *Age*, and *Sex*, and their relationship between other clinical information. We also look into the multicollinearity between the variables to find potential interaction terms.

Distributions

We convert the response variable into binary based on the positiveness of its value, although we find that there are much more negative values than positive ones (2:1) (On the left of Figure 1), such imbalance ratio is tolerable. The distribution of race is also disproportional, with 81.21% of the people in the training dataset are White, 17.75% of them are other, and 1.04% of them are Black. The distribution of *Sex* is relatively more balanced with female taking up 60%.

Grouping *Age* to 3 categories including *Young*, *Middle Age*, and *Elder* based on their medical definitions (cite), we find that for all race groups, young people have the highest proportion. *White* group has a significantly higher percentage of young people, and relatively similar middle aged and elderly people. For *Black* and *Other* racial groups, the differences among the number of young, middle aged, and elderly people are less significant. *Female* group also has a significantly high percentage of young people. The numbers of observations for middle aged female, elderly female, young male, middle aged male, and elderly male are relatively similar.

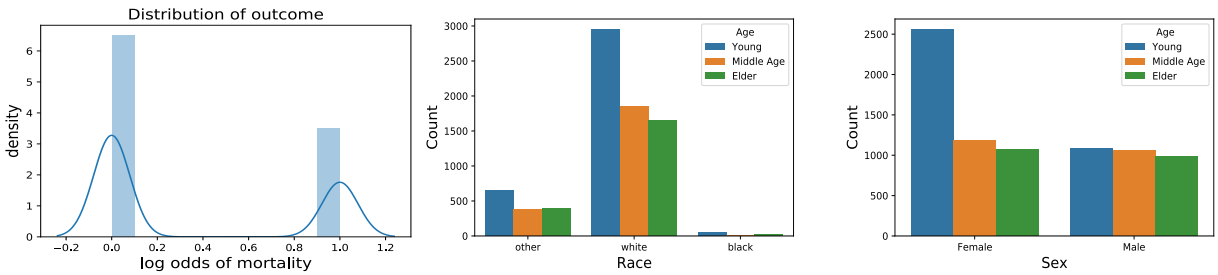


Figure 1: Distribution of converted output & Counts plot of sex and race, color-coded by age

Relationship Between *Age-Sex* and Other Predictors

We cut the data once more based on both *Age* and *Sex* to create 6 distinct *Age-Sex* groups, for example *Young female*, solely for EDA purposes. In Figure 2 below we present the relationship of 3 predictors, clinical and socioeconomic, between *Age-Sex*.

Based on *Poverty index* vs. *Age-Sex*, we observe that People with high poverty index tend to die before the end of the study (*post-study mortality = 0*). This trend is especially obvious for both non-elderly male and female. Based on *BMI* vs *Age sex*, we observe that female of all age groups who die after the end of study (*post-study mortality = 1*) have a higher *BMI*. This trend does not apply to male as much as to female. Based on *Systolcic BP* vs *Age sex*, we observe that those who die after the end of study (*post-study mortality = 1*) tend to have higher *Systolcic BP*. We also observe some differences in the distribution of *Systolcic BP* across age groups. For example, middle aged females and elderly males have a nearly normal distribution. Young females, young males, middle aged males, and elderly males have more than 1 peaks for those who die before the end of study (*post-study mortality = 0*).

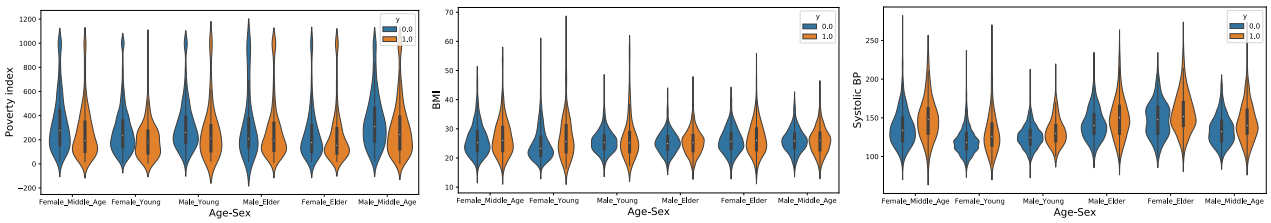


Figure 2: Violin plot of Poverty Index, BMI, Systolic BP (left to right) vs. Age-Sex

Multicollinearity

We use the raw data which consists of mostly numerical variables to construct a correlation matrix. The heat map in Figure 3 shows that multicollinearity does exist. Pairs of predictor that have high correlations include *TS* and *Serum iron*, *Pulse pressure* and *Systolic BP (SBP)*, *SBP* and *Diastolic BP (DBP)*. This observation makes sense because *TS* is calculated from *Serum iron*, *Pulse pressure* is calculated from *SBP* and *DBP*.

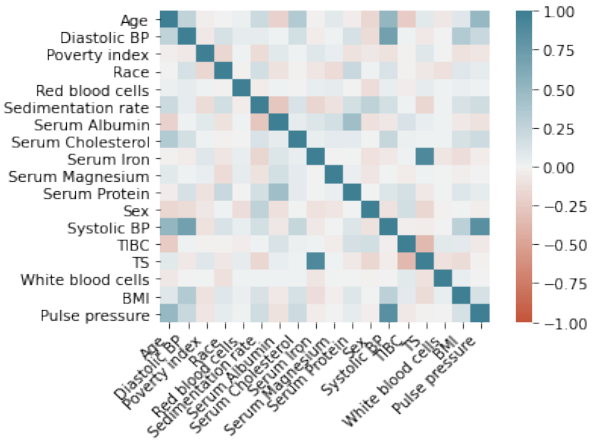


Figure 3: Correlation Matrix

3 Models

3.1 Modeling

To tackle the project question, we start from baseline models such as logistics regression and decision, gradually adding regularization and boosting techniques. Our modeling process includes the comparison of models from largely 6 types: logistic regressions, decision trees, random forest, gradient boosting decision tree (GBDT), AdaBoost and long short-term memory (LSTM) network. We are inspired by literature [1], and our addition is mainly in the gradient boosting part using the LightGBM framework and the utilization fo LSTM network.

3.1.1 Logistic Regression

Logistic regression is a regression method for a binary outcome, which applies to our problem when using binary *post-study mortality* as our response variable. Logistic regression models the log odds of a person to live after the end of the study.

We construct 5 different logistic regression models. For *Model 1*, we use a default logistic regression, then we add ridge-like regularization to our *Model 2*. The bpenalty term is added to the model to penalize large coefficient values to resolve potential overfitting issues. Further, we tried adding interaction terms in *Model 3* and polynomial terms to the maximum degree of 3 in *Model 4*. Finally for *Model 5* , we build on *Model 4* by adding lasso-like regualrization (L1), which means that not only a penalty is added to penalize large coefficient values, insignificant variables will be eliminated by setting their coefficients to 0. L1 regularization reduces the number of features we have.

3.1.2 Decision Tree

For *Model 6*, we build a simple decision tree with maximum depth of 3. In a decision tree classifier, we first split on the feature in the root node, which is usually one with high feature importance. We then split another time on the 2 partitions resulted from the previous split. When we stop split, partitions that is left are leaves, and each leaf is classified to the class that is in majority in that leaf. One advantage of a decision tree classifier is that it is easy to interpret and extract feature importance.

3.1.3 Gradient Boosting Decision Tree

For *Model 7*, we build a GBDT model using the LightGBM framework. This frame work is high in training speed and efficiency. GBDT is an ensemble tree building method. We start with building a decision tree, and we use the loss of this prediction to update the prediction of the next tree. This means that a subsequent tree learns from the previous mistakes, and the previously misclassified predictions are given higher weights to be learned.

The set of hyper-parameters we use includes *feature_fraction=0.9*, *bagging_fraction=0.9*, *bagging_freq=5*. This means that for each iteration where a tree is built, LightGBM randomly selects 90% of the features to include in the model. For every 5 iterations, we perform bagging by bootstrapping 10% of the data and leave 90% of the data not resampled. Both of these methods can effectively reduce overfitting and help the model be more robust. Therefore, we use GBDT as our model for feature selection because it gives a more averaged model that reflects the true effect of the resulting feature space after dropping a variable.

One advantage of GBDT using the LightGBM framework is that the package easy to implement replicate for future improvement and applications. Also, same as decision tree, it is also easy for GBDT to produce feature importance. We use GBDT to produce SHAP values which are used for feature selection in the subsequent section.

3.1.4 Random Forest

For *Model 8*, we build a random forest model. Random forest is a method that builds multiple decision trees with random subsets of features. We start with training each tree on a separate bootstrap sample of our data set. When building each tree, the feature used is randomly selected. Then we average the outcomes of all trees to get the random forest prediction. Randomly selecting the set of features can reduce the effect of dominating features always being split at early stages.

3.1.5 AdaBoost

For *Model 9* we AdaBoost, another ensemble method. We start with a base estimator. In our case, we use a decision tree classifier with maximum depth of 3 (same estimator as *Model 6*). We initialize the weight of our training data to be $\frac{1}{\text{number of training points}}$, fit a tree on it, and update the weights based on the prediction of each data point: misclassified points will have larger weights. We then fit another tree on the weighted training data and update the prediction on top of the previous trees iteratively.

3.1.6 Long Short-Term Memory Network

LSTM networks are a type of Recurrent Neural Network (RNN). Each batch of data is fed into the hidden layers of the network sequentially so that the output of the previous epoch will be the input for the future epoch. For *Model 10*, we build the LSTM using the package *Keras*. Our architecture consists of LSTM layers of 128, 64,

32, and 16 units, all with *Relu* activation, followed by an output layer of 1 unit with *Sigmoid* activation. We use Adam optimizer and add early stopping criteria to reduce potential overfitting.

LSTM is particularly different from other standard RNNs because in standard RNNs, the gradient of the loss function decays exponentially as we run more epochs. LSTM networks have special memory cells that enables them to retain information and continuing learning from them for a longer period of time. Another advantage is that using this package is easy to implement and replicate for future improvement and applications.

3.2 Model Result Comparison

Table 3 and the ROC graph (Figure 4) show a comparison of the performance among all 10 models we build in this project. The ROC graph illustrates the diagnostic ability of our models. Notice that the area under the curve is the AUC. The closer to the upper left corner the line is (the larger AUC), the higher the tree positive rate over the false positive rate will be. We can see that models such as LSTM, GBDT, random forest, and logistics interaction fall together with the largest AUC, but models such as single decision tree, logistics with polynomial terms, and AdaBoost are associated with smaller AUC.

From the table we have the exact numerical comparison of test accuracy and AUC. In terms of test accuracy, GBDT and LSTM perform equally well, reaching a testing accuracy of 84.15% and 83.74%, followed by random forest with 83.59%. In terms of test AUC, GBDT performed the best with 81.99%, followed by LSTM with 81.85% and random forest with 81.80%. In terms of test AUC, GBDT performed the best with 81.99%, followed by random forest with 81.80% and LSTM with 81.34%.

We do not observe overfitting issues for logistic regressions *Model 0 - 3*, but overfitting occurs in *Model 0 & 3*, where polynomial terms are added. For these 2 models, training accuracy is successfully improved from about 83% to about 87%, but testing accuracy decreased compared to logistic regressions without polynomial terms.

Results of random forest and AdaBoost show that both models have overfitting issues. The training accuracy of random forest model reaches 90.93%, and training accuracy of AdaBoost even reaches 95.63%. Their testing accuracy, as shown in Table 3, remain much lower than the training accuracy. Notably, GBDT performs relatively well with not only the highest testing accuracy, but also minimal overfitting, with training accuracy being 86.23%, only slightly higher than testing accuracy which is 84.15%.

Table 3: Results

Model Type	Model Details	Test AUC	Test Accuracy
Logistic regression	Default	81.29%	83.49%
	L2	81.33%	83.54%
	Interaction	81.42%	83.49%
	Polynomial	78.36%	80.67%
	Polynomial & L1	79.17%	81.38%
Decision Tree	Max depth = 3	81.03%	82.64%
GBDT (LightGBM)	Max depth = 100	81.99%	84.15%
Random Forest	n_estimators = 200	81.80%	83.59%
AdaBoost	Base: decision tree with max depth = 3	78.40%	81.03%
LSTM	# units in each layer: 128, 64, 32, 16	81.34%	82.73%

Note: Model details contain selected notable parameters. They are not all of the parameters in the corresponding model.

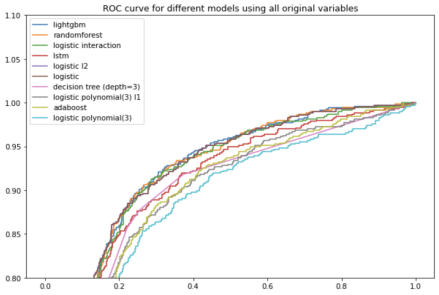


Figure 4: ROC curve for different models using all original variables

3.3 SHAP Interpretation

As we recognize that GBDT model gives the best performance, we turn our attention to the interpretation of each feature's effect in the model. By combining many local explanations we can provide rich summaries of both an entire model and individual features. Figure 5 is the SHapley Additive exPlanations (SHAP) summary plot for GBDT model trained on the mortality dataset. Shapley values tell us how to fairly attribute the prediction among the features [7] (but not necessarily the causation). We can summarize that the primary risk factor for death according to the model is how aged one is.

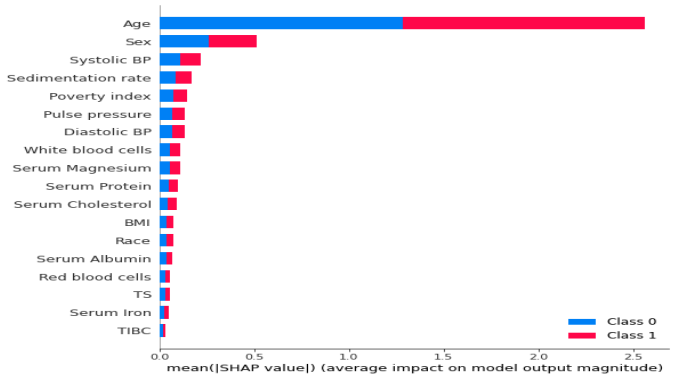
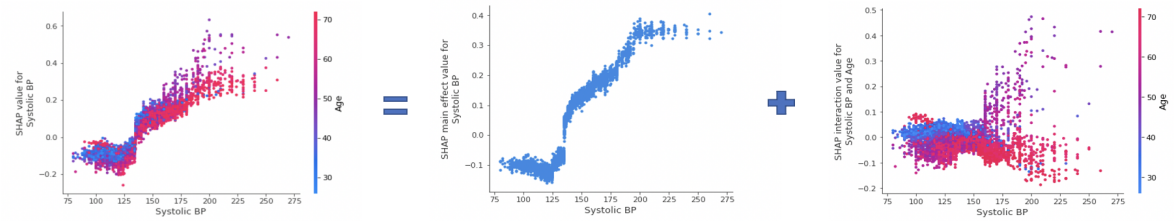


Figure 5: SHAP summary plot

We choose the top 4 predictors (Age, Sex, Systolic BP, Sedimentation rate) form the summary plot to construct SHAP dependence plots Figure 6 to illustrate how each feature affects the outcome of the model.



(1) SHAP dependence plot of Systolic BP vs. its SHAP value

(2) remove the interaction effect of age

(3) the interaction effect of systolic blood pressure with age

Figure 6: SHAP dependence plot for systolic blood pressure and age

Using SHAP interaction values we can remove the interaction effect of age from the model(Lundberg, 2019 [6]). Figure 6 (1) shows the overall effect of systolic blood pressure on the model. It is expected that when the systolic blood pressure goes up, the risk of dying increases. However, the effect varies with age and can be decomposed. Figure 6 (2) shows the single effect of systolic blood pressure on predicting the post-study mortality, the higher the systolic blood pressure, the higher post-study mortality as our model tells. In figure 6 (3), plotting just the interaction effect of systolic blood pressure with age shows how the effect of systolic blood pressure on mortality risk varies with age.

Figure 7 shows the selected overall effect and interaction effect of remaining pairs. In (1), the post-study mortality rate increases as the sedimentation rate increases, and the effect is more obvious for female when the sedimentation rate is high. It aligns with the finding [10] that women demonstrated average sedimentation rate higher than male did, and therefore in the higher sedimentation rate regime, female tends to have higher mortality rate. (2) and (3) illustrate the effect of systolic BP colored by Sex and Sedimentation rate. When systolic BP is considered as high (around 140 [12]), female is more susceptible to mortality. Nevertheless, as systolic BP becomes higher, the sex interaction diminishes. The sedimentation rate has less obvious interaction effect on systolic BP, and the trend and strength when sedimentation rate is used to predict the mortality is similar to those of systolic BP.

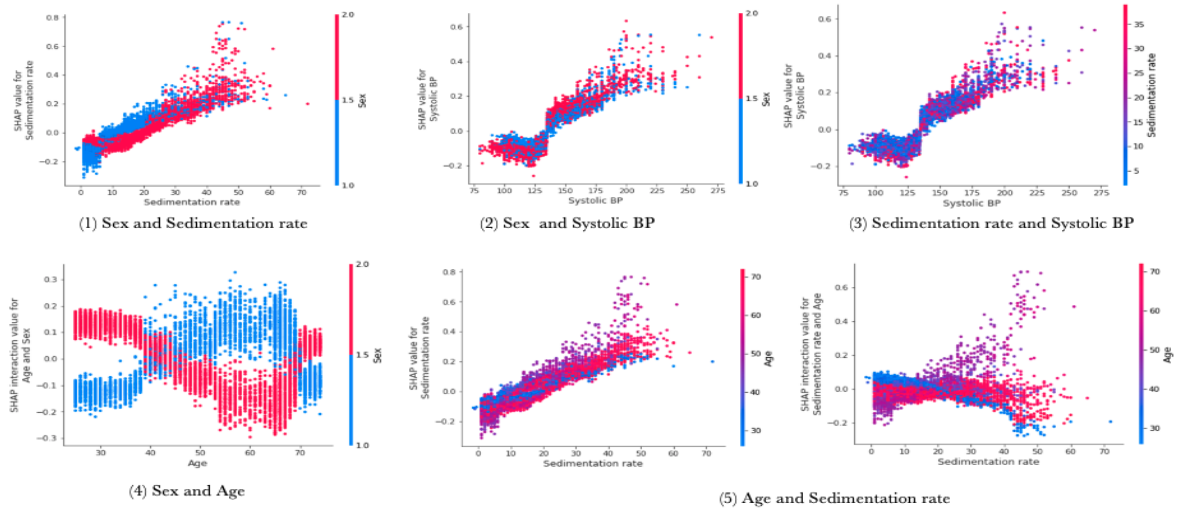


Figure 7: SHAP dependence plot for remaining pairs

Figure 7 (4) of age and sex shows a clear flip in the relative risk between men and women over a lifetime. When the age is near 60, female and male has largest difference in mortality rate. As suggested in the literature(Lundberg, 2019 [6]), it is plausible that this increased risk is driven by increased cardiovascular mortality in men relative to women near that age. In the right subplot of (5), the interaction effect gives a closer look besides the entanglement of age's effect on sedimentation rate. For older and younger people, the sedimentation rate has a stable effect across the range. However, for mid-aged people, the higher the sedimentation rate, the higher probability of mortality.

4 Feature Selection

In our modeling process, we observed that having a single *Age* predictor will result in around XX% accuracy in GBDT model, and 82.64% in Logistic Regression. But having only one predictor in the model will become vulnerable to the extreme values of that predictor, and the accuracy is not stable and robust across different models. From feature importance section above, the SHAP interaction plots clearly show the interaction effect do exist between *Age* and other variable, using *Age* only will also lose important information. For the sake of making our model more robust, we conduct the feature selection.

We aim to choose the smallest set of features which makes our model still robust and accurate. Considering we only have this data and we hope to have a broader application of our model, we combine the technical results and insight from literature to reach our conclusion. We start with our GBDT Model using LightGBM framework to do the feature selection. From the feature importance result we obtained from the modeling section (Figure 5, we can clearly see that *Age* has the highest SHAP value, followed by *Sex*, *Systolic BP*, *Sedimentation rate*, *Pulse Pressure*, *Poverty Index*, and *Serum Cholesterol*.

The model gives a scope of top variables that we can put into our final subset of predictors. However, since *Age* is such an strong predictor, including any subset of other top variables is not significantly varying the model result. Then, we turn our attention to the literature.

Not surprisingly, the top variables from feature importance plot overlap with some of the literature: *Systolic BP* [12], *Sedimentation rate* [13],*Pulse Pressure* [2], and *Serum Cholesterol* [4] are significant predictors in mortality rate.

To figure out the best quantity and combination that we will use, we design a relatively automatic way of selection. First, we fit a GBDT model in LightGBM with with 90% feature included each time, for a bagging frequency 5 times and drop the variable with the least SHAP value. Then We iterate the logic until all 18 variables have been dropped, while recording the testing and evaluation accuracy. In the meantime, we fit a logistics regression with L2 regularization and record the testing and evaluation accuracy along the way, when the variable is dropped in each step as a benchmark comparison.

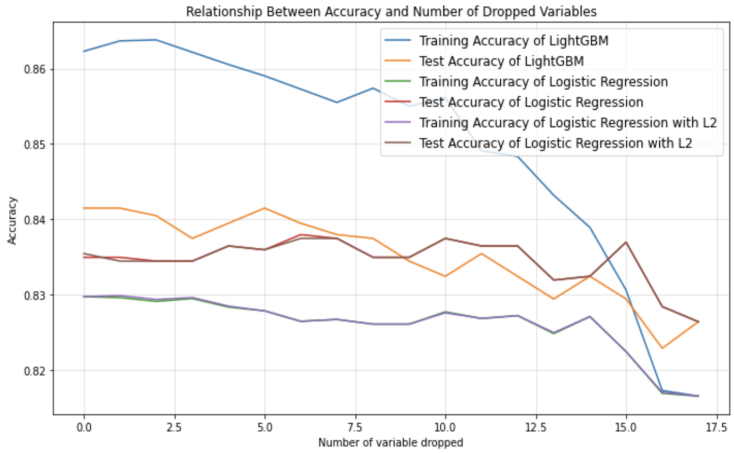


Figure 8: Feature Selection Process

The logistic regression serves as a baseline reference that guides us to the stopping point of dropping. The horizontal label denotes the number of variable dropped. We notice that the training and test accuracy for GBDT model drops as the number of variable dropped increases. When the number of variable dropped in less than 14, the training and test accuracy for logistic with L2 stays stable. However, when there are 15 variables dropped, the logistic with L2's training accuracy steps down from the plateau. Moreover, the test accuracy of GBDT drops below 83%. Hence, we decide to stop the dropping when 14 variables are dropped, leaving 4 predictors in the reduced model.

We records each variable's frequency of ranking top 4 in SHAP values during the feature selection process. Due to the slight randomness of GBDT model, the result we get each time is slightly different. In general, predictors with the greatest frequencies are *Age*, *Sex*, *Systolic BP*and *Sedimentation rate*. We used the model to get a test accuracy of 83%, which is comparable as we have used all 18 predictors. The result coincides again with our investigation on the literature. It follows the common sense that the older the person, the higher risk of mortality it presents. Furthermore, as *Sex* is coded as 1: Male, 2:Female, the probability of mortality is greatly influence by the gender. The high systolic BP is related to risk of strokes, corona heart disease and chronic kidney disease, and the low systolic BP will cause other problems (cite). CHD and kidney disease are often time fatal or have long impact even after medical treatment, so they are significantly associates with the risk of death. The same reasoning applies to the sedimentation rate. It is related to different blood disease and disorder in the immune system, which is understandably strongly related to the mortality of human being.

5 Robustness in modeling

Throughout the modeling process, we make two crucial decision on the independent variables and the dependent variable. We decide to use the independent variables directly from the original dataset, without the grouping transformation mentioned in the data handling section. With the original independent variables our best modeling result achieves 84.2% test accuracy, and the test AUC of 83.24% the highest accuracy and highest AUC.

We have tried to use the categorical independent variables as we mentioned in the data handling section. However, since *Age* is a very strong predictor, we have not improved or our accuracy using the grouped independent variables. We get the best test accuracy to be 82.89% using the GBDT model using LightGBM framework, and we still have categories of *Age* to be the predictors that have most impact on model shown in the left subplot of Figure 9. In addition, creating dummies for each category of each categorized variable result in a significant increase in the number of features, making it harder to interpret. Therefore, we consider modeling with original, non-transformed independent variables.

We also decide to use binary output variable (our post-study mortality) instead of a continuous, original response variable. We compared our modeling result with the model with continuous output, and we find out the predictions of the post-study mortality is deficient in accuracy. As shown in the right sub plot of Figure 9, blue-colored predicted values fall around zero, however the true value clustered away from zero. The continuous prediction is greatly affected by the negative values, and different types of models inevitably have large MSEs. That justifies that our model with binary output (post-study mortality) is effective.

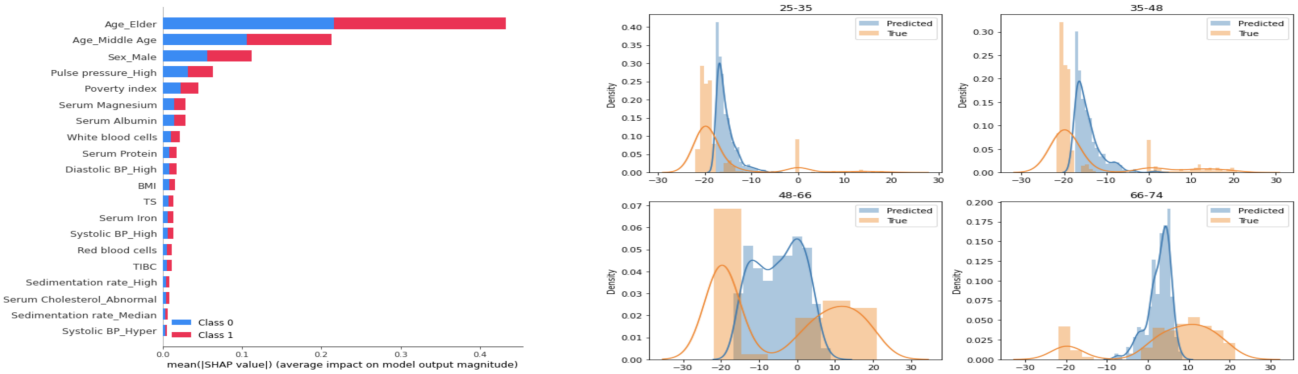


Figure 9: SHAP Summary Using Categorized Input & Distribution of True and Predicted Using Continuous output

6 Conclusion

6.1 Conclusion

In conclusion, we have achieved several goals through this research project. First, enlightened by the literature, we build several models from the simplest logistic regression to GBDT and LSTM. In terms of predicting the mortality rate using the data set given, we discover that the GBDT model with LightGBM framework produces the best prediction accuracy, best Area Under Curve, and minimal over-fitting issues. Second, we are able to use SHAP values and dependence plots to identify the effect of top features on mortality rate on our GBDT model. Then, we manage to find the smallest subset of features that will give us comparative model performance using an innovative approach. That is, we combine the high standard of statistical metrics (accuracy) with the implication from medical literature, and justify that age, sex, systolic blood pressure, and the sedimentation rate are the most important factors affecting the risk of dying. What's more, we have justified our modeling choices of binary dependent variable and original, un-grouped independent variables. Having a binary output variable minimizes the risk of underestimating the mortality rate of young people, and having continuous output makes our model more concise and precise.

Our finding confirms partly the finding of the medical community that factors such as age, sex, systolic blood pressure, and sedimentation rate are highly associated with the mortality. Moreover, it probably gives physical health-related suggestion for the general public.

6.2 Limitation and Future Improvement

There are certain limitation and aspects of future improvements. To start with, our data comes from a survey dated back to 1992 and the we only have access to the subset of data. The result of modeling on earlier data set could give a general implication on mortality rate prediction, but in order to get more specific insights, we could try to get a more recent data to work with. Furthermore, if we could obtain the data across different years, we would be able to conduct a more throughout analysis using time series prediction or survival analysis, as some medical papers suggest.

There might be a better way to use continuous (original) output in our model. We choose in this project to convert continuous output into binary because of the interpretation of negative y values is unclear. If we have more information and understandings of each feature and the output variable, we can do better in data handling.

Our LSTM model has room for improvements, too. In the future, we could try on package such as $PyTorch$ other than our current usage of $tensorflow.keras$. The former packages is supposed to give users more flexibility in defining the parameter and other functionalities from experiences. That adjustment might lead to a better model performance.

7 Broader Impact and Inclusivity Statement

7.1 Broader Impact

7.1.1 Benefits

Predicting the likelihood of patients' mortality as a binary outcome is an important area in biomedical research and clinical practices. Built upon the gradient boosting ensemble model – LightGBM which yields an accurate prediction of the probability of mortality, our team also proposed a feature selection framework which could be broadcast to figure out the most significant variables on predicting mortality outcome, generating broader applications in clinical trials, healthcare consulting and medical research.

7.1.2 Risks

Ethical Issue: It turns out gender is an important variable. Although it is well-known that mortality rate and the incidence of diseases varies across gender, the machine learning model and the interpretation should be gender-agnostic. Race and Poverty Index can also be involved in ethical problems. Although they are not among the 4 important features, we still need to stress on the point that our research treats every race, culture, gender, class, religion, etc equally.

Data Integrity Issue: We put data privacy, security and integrity at the top of our priority. Our dataset is real-world data, coming from long-term surveys of large-scale research. While processing past dataset or future dataset, we strongly encourage to maintain the accuracy of patients health summary, demographic information, medical record. Any activity that breaches data security can lead to a corruption of our effort and run counter to our research purpose

7.2 Inclusivity Statement

First, the purpose of data science is to discover the hidden pattern from the ocean of data, build models for accurate and efficient prediction and interpret and recommend the conclusion for social good. The conclusion we came up with is only based on NHANES-I dataset. We discourage any illegal usage of our data and model, any misconduct of the results, any discrimination toward different groups, any disparity in healthcare coverage and any threats in national security.

Second, we hope to serve as a contributor to boosting inclusivity and engaging the participation of minority groups (female, economically disadvantaged groups, underrepresented groups, etc) in data science. We look forward to receiving any criticism or correction upon our methodology, and any suggestions on future directions from any institution or individual!

Acknowledgement

We kindly acknowledge the guidance and assistance we received from the CS109 team.

References

[1] Austin, Peter C, Lee, Douglas S, Steyerberg, Ewout W, & Tu, Jack V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biometrical Journal, 54*(5), 657-673.

[2] Benetos, Athanase, Safar, Michel, Rudnichi, Annie, Smulyan, Harold, Richard, Jacques-Lucien, Ducimetiere, Pierre, & Guize, Louis. (1997). Pulse Pressure: A Predictor of Long-term Cardiovascular Mortality in a French Male Population. *Hypertension (Dallas, Tex. 1979), 30*(6), 1410-1415.

[3] Carson, Jeffrey L, Dey, Achintan, Milan, Edwin, Russell, Louise B, Taylor, William C, & Jagannathan, Radha. (1998). Modeling all-cause mortality: Projections of the impact of smoking cessation based on the NHEFS. (National Health and Nutrition Examination Survey Epidemiological Followup Survey). *American Journal of Public Health (1971), 88*(4), 630.

[4] Garg, Amit X, Clark, William F, Haynes, R. Brian, & House, Andrew A. (2002). Moderate renal insufficiency and the risk of cardiovascular mortality: Results from the NHANES I. *Kidney International, 61*(4), 1486-1494.

[5] Greenberg, J. (2001). Biases in the mortality risk versus body mass index relationship in the NHANES-1 Epidemiologic Follow-up Study. *International Journal of Obesity, 25*(7), 1071-1078.

[6] Lundberg, Scott M, Erion, Gabriel, Chen, Hugh, DeGrave, Alex, Prutkin, Jordan M, Nair, Bala, . . . Lee, Su-In. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding.

[7] Molnar, Christoph. 5.10 SHAP (SHapley Additive ExPlanations) — Interpretable Machine Learning. christophm.github.io, https://christophm.github.io/interpretable-ml-book/shap.html. Accessed 12 Dec. 2020.

[8] NHANES I - Epidemiologic Followup Study (NHEFS). https://wwwn.cdc.gov/nchs/nhanes/nhefs/default.aspx/#pudf. Accessed 12 Dec. 2020.

[9] Sharma, Praveen, Dietrich, Thomas, Ferro, Charles J, Cockwell, Paul, & Chapple, Iain L.C. (2016). Association between periodontitis and mortality in stages 3-5 chronic kidney disease: NHANES III and linked mortality study. *Journal of Clinical Periodontology, 43*(2), 104-113.

[10] Siemons, Liseth, Ten Klooster, Peter M, Vonkeman, Harald E, Van Riel, Piet LCM, Glas, Cees AW, & Van de Laar, Mart AFJ. (2014). How age and sex affect the erythrocyte sedimentation rate and C-reactive protein in early rheumatoid arthritis. *BMC Musculoskeletal Disorders, 15*(1), 368.

[11] Thuluvath, Paul J, Yoo, Hwan Y, & Thompson, Richard E. (2003). A model to predict survival at one month, one year, and five years after liver transplantation based on pretransplant clinical characteristics. *Liver Transplantation, 9*(5), 527-532.

[12] "Understanding Blood Pressure Readings." *Www.Heart.Org*, https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings. Accessed 12 Dec. 2020.

[13] Veeranna, Vikas, Zalawadiya, Sandip K, Panaich, Sidakpal, Patel, Kushang V, & Afonso, Luis. (2013). Comparative analysis of red cell distribution width and high sensitivity C-reactive protein for coronary heart disease mortality prediction in multi-ethnic population: Findings from the 1999–2004 NHANES. *International Journal of Cardiology, 168*(6), 5156-5161.