Mortality Prediction and Interpretation

Li Sun, Yanqi Luo, Hao Ming Chen, Shengnan Xu

1. Introduction and Motivation

The prediction of mortality based on social, clinical, nutritional and behavioral features serve a multitude of useful social and commercial functions. Insurance companies and medical care administrations, for example, have an interest in understanding the mortality risks of patients in order to identify those in most urgent needs of medical and clinical resources as well as conserve limited healthcare resources by de-prioritizing those in a low risk group. Mortality predictions also serve key functions for governments and medical researchers interested in devising the best public health strategy to promote citizens' well-being. For example, with a limited budget, what aspects of behavioral or nutritional health should they emphasize that best decrease citizens' risk of dying? To these ends, we ask the following two questions in our project: which model would best predict a person's risk of dying? How do we interpret the impact of predictors and what are the most important features for our model?

2. Data

2.1 Data Source and Data Description

The dataset we used in this project was collected as part of the National Health And Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS), which as a longitudinal study covering the national population initiated by the National Institute on Aging and National Center for Health Statistics. The goal of the NHEFS study was to investigate the relationships between clinical, nutritional, physical and behavioral attributes and mortality, morbidity, hospital utilization, and impact on risk factors. In this dataset, we investigated the relationships between many of the clinical, physical, nutritional, and social variables measured in the National Health And Nutrition Examination Survey I and subsequent mortality of the traced cohort. There were 46 variables, including 41 predictors and 1 response variable. The response variable represented the time the study participant had survived after the first examination. The value for this variable was negative if the participant had dropped out of the study prior to completion or was still alive at completion. Among these 41 predictors, 3 (urine_albumin, urine_glucose, urine_hematest) were ordinal, 3 (race, sex, and platelets_estimate) were categorical and 35 were continuous. In order to use these categorical and ordinal variables in modeling, we created dummy variables for all categorical variables using one-hot encoding and converted descriptions of levels for all ordinal variables on an integer scale from 1 to 6.

2.2 Data Reconciliation

In order to confirm validity of data, we performed data reconciliation. To gauge the high-level representativeness and validity of participant data collected, we cross-referenced values of common and well-documented physiological measures (red and white blood cell counts, and hemoglobin and hematocrit levels) in our dataset against reference ranges (Figure 3, Table 1). Using reference ranges from the Mayo Clinic [1], a premier healthcare provider, we saw that values for all variables tracked closely the reference values for both males and females. For example, the red blood cell counts mostly fell between 4.2 and 5.8, as well as 3.8 and 5.25 for males and females in our dataset whereas the reference ranges for males and females (adults) were 4.35-5.65 and

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

3.92-5.13, respectively. Furthermore, for variables such as red blood cell count and hemoglobin where reference ranges were different for males and females, a similar shift in distribution of values was observed for these variables in our dataset. Taken together, these observations supported the conclusion that the cohort in our dataset was representative of the adult population of this country and that there were no significant validity issues regarding the data used.

2.3 Missing Value Imputation

Due to significant data missingness in the dataset, we conducted data imputation and variable selection prior to modeling and data visualizations (Figure 1). We first visualized and quantified the missingness of variables. Together, we saw that 15 variables were missing more than half of the values and 4 variables were missing more than 75% (Figure 2). Because variables missing more than 75% values were unlikely to carry any beneficial predictive value and could even be detrimental due to highly speculative imputation, we dropped them from the dataset. For the remaining 11 predictors missing more than half of the values, we converted them to indicators of missingness (1=variable present, 0=variable missing) in order to balance the need to have more variables as predictors and the potential hazard from imputation of variables that did not contain sufficient information. Because our EDA showed that there were no significant linear relationships among the variables, we decided to apply kNN instead of linear regression imputation. After applying the minmax scaler, We applied two rounds of kNN imputation. In the first round, three variables with no missing values were used to impute variables with less than 5% of missing values. In the second round, variables with between 5% and 50% missing values rate were imputed based on variables with non-missing values and variables missing less than 5% that had been imputed in the first round.

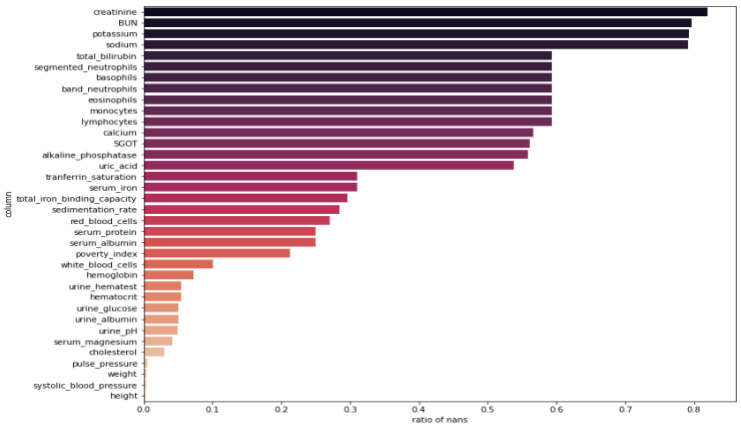

Fig 2. missing values ratios in each predictors with missing values

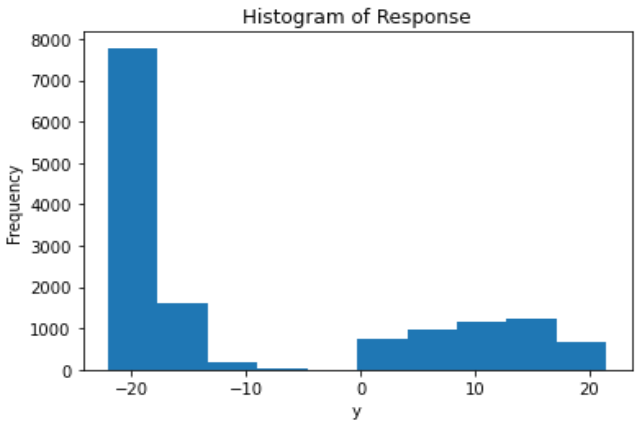

Fig 4. Distribution of the response variable

3. Exploratory Data Analysis

We visualized the distribution of the response variable as well as predictors against the response variable to explore the possibility of utilizing linear models for prediction. However, no obvious linear relationships were observed between continuous predictors and the response variable (Figure 6). In addition, the response variable did not appear to be normally distributed (Figure 4), which would violate assumptions for linear models such as linear regression.

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

We also sought to explore multicollinearity in our dataset by visualizing a correlation matrix of continuous predictors against each other (Figure 5). We observed possible issues with multicollinearity as variables had modest to strong correlations amongst each other. For example, pulse pressure had 0.5 and 0.85 correlation coefficients with age and systolic blood pressure, respectively (Figure 5).

We then moved on to consider any interaction effects that might be present, especially regarding race and gender. Looking at the effect on the response variable of platelets estimates across genders and of race across genders, for example, we saw that race and platelets estimates might have been modified by gender (Figure 8-9).

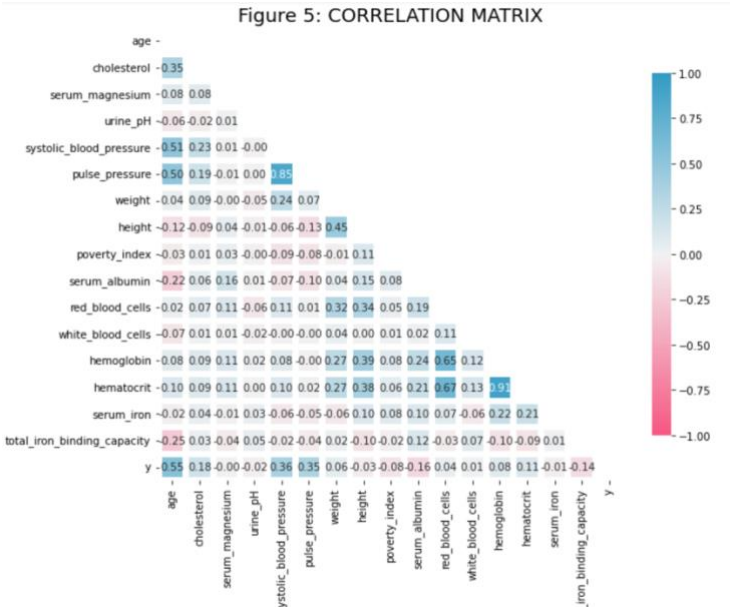

Fig 5. Correlation Matrix of Variables

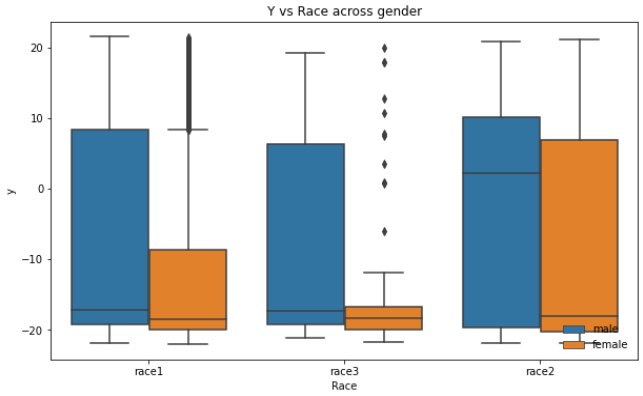

Fig 8. Distribution of response variable in different race and gender groups

Finally, we examined the relationship between categorical variables (Gender, PE, and race) and the outcome variable to determine their usefulness in predicting the response variable (Figure 10-12). We used histograms together with violin plots to compare and distributions of target variables within each subcategory. We could see that the distribution of y was inconsistent across subgroups. Therefore, it was worthwhile to incorporate categorical variables into the final prediction model.

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

Because the response variable was not distributed normally (Figure 4) and because we sought to answer a classification instead of regression problem, we converted the continuous response variable into a categorical one based on whether the participant had survived x number of years. Taking this approach, we would delete negative y values whose absolute values were smaller than the cut-off x because we had no way of knowing if participants with those values died during the study or dropped out. We experimented with different values for x and visualized the proportionality of classes (alive vs. deceased) with different thresholds values (Figure 13). We eventually selected a threshold of 19 because this would leave us with the most equal distribution of participants who survived or did not survive past that number of years since their initial examination (Figure 13). We then took the absolute value of the response variable and created a new variable alive as our outcome variable where participants would have values of 1 for that variable if they survived at least 19 years since the first examination and 0 if they did not.

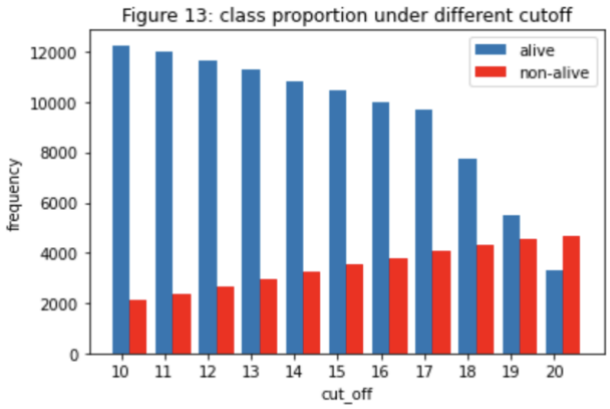

Fig 13. Class Proportion under Different Cutoff

After transformation of this response variable, we conducted further visualizations to analyze the difference in values in different predictors between alive and deceased patients using the selected threshold (Figure 14). Using KDE and bar plots, we noted that, not surprisingly, age and gender appeared to have substantial effects on participants' mortality, with older people and males more likely to die in 19 years since first examination of the study. We also noted some interesting findings: for example, poverty, pulse pressure, systolic blood pressure, and height seemed to have relatively significant impact on mortality while weight, white blood cells, and serum iron appear to have very modest if any impact at all (Figure 14).

4. Methods

4.1 Advanced Feature Engineering and Selection

4.1.1 Feature Selection

We used two methods to conduct the feature selection. The first method was lasso regularization because it had the tendency of "turning off" unimportant predictors. Specifically, we removed all predictors that the model deemed unimportant. The second method was random forest. After fitting a random forest model, we calculated the feature importance for each predictor. A threshold

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

(0.007) was set to remove all predictors that had scores lower than it. Only the features that were kept both in the lasso process and the random forest process would be kept in our feature subset.

4.1.2 PCA Transformation

PCA has the unique ability to effectively deal with problems of multicollinearity as well as to reduce runtime and complexity through reducing the dimension of predictors. Therefore, in addition to using the original set of predictors selected only based on data missingness (original predictors) and the aforementioned subset of predictors selected through application of lasso regularization and random forest (subsetted predictors) as input data to our model, we also sought to consider PCA transformed original and subsetted predictors to our prediction model. To this end, we performed PCA transformation of these two sets of predictors. In determining the number of principal components needed for each set of predictors, we aimed for criteria such that 85-90% of variance could be explained.

4.2 Modeling

After the aforementioned process of feature engineering, we created four different datasets: original dataset, subset using feature selection only, dataset using PCA only, subset using both feature selection and PCA. We wanted to explore whether PCA could reduce model complexity while improving the model's performance. We also wanted to explore whether we could use a subset of features to achieve similar performance as the original dataset, which could bring great convenience in clinical practice.

First we employed K-nearest neighbors as our baseline model on the original dataset since it did not require normality, linearity and multicollinearity assumptions in data and was more capable of building complex decision boundaries automatically to offer useful predictions. Then, logistic regression, decision trees, random forest, adaboosting, SVM, neural network were employed on the original dataset to try to improve the performance. We decided to choose the two best performing models to further implement on the 3 other dataset.

4.2.1 K-Nearest Neighbors - Baseline Model

K-nearest neighbors (KNN) has been viewed as one of the top 10 data mining algorithms due to its efficiency and simplicity [2]. It is a model that performs classification by first calculating the distance between the observant and every training sample, then returning the mode of the k nearest samples' labels. In order to select the best hyper-parameter k for our kNN model, we employed 5-fold cross validation on our train data on a range of k values. We selected the best k value based on the training accuracy and validation accuracy (Figure 19).

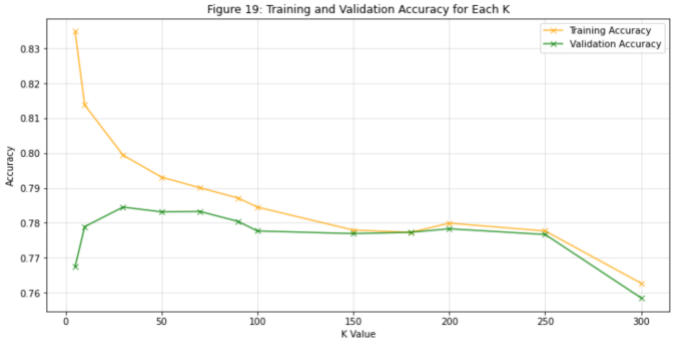

Fig 19. Training and validation accuracy in KNN model for different k values

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

4.2.2 Logistic Regression

Logistic regression is one of the most common generalized linear regression, which is frequently used for binary classification. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution.

$$logit\big(\mathrm{P}(\mathrm{Y}=1)\big) = \log\frac{\mathrm{P}(\mathrm{Y}=1)}{1-\mathrm{P}(\mathrm{Y}=1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k,$$

$$\mathrm{P}(\mathrm{Y}=1) = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\beta_k x_k)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\beta_k x_k)},$$

4.2.3 Decision Trees

Decision tree is a tree-structured model based on a series of comparisons of the values of predictors against threshold values. It is easy to interpret, and could build complex decision boundaries automatically. We tuned the number of max_depth of trees through cross-validation to find a well-tuned single tree.

4.2.4 Random Forests

Random forest addresses the high variance issues arising from the deep depth of a single decision tree by combining the prediction results from multiple decision trees to reduce the variance. We tuned the number of trees, max_depth of each tree, and loss function in determining splits to select the optimal combination of hyperparameters on the average result of 5-fold cross validation.

4.2.5 SVM

Support Vector Machine takes advantage of the kernel function to project the original data points into a higher dimensional space or even infinite dimensional space with the RBF kernel. The linear decision boundary in higher dimensional setting will then become a non-linear decision boundary in the original dimensional space. Similarly, we employed a grid search and only varied the kernel function and regularization loss weight in the hinge loss.

4.2.6 Neural Network

We defined a simple dense neural network with three hidden layers and one output layer. To capture the non-linear trend, we added a ReLU activation function after each hidden layer. In the output layer, we used a sigmoid activation function to convert the raw number to probability and calculate the loss with binary cross-entropy functions. We trained the neural network with 100 epochs with EarlyStopping callbacks to stop training when there was no improvement in the validation accuracy. We manually tuned the learning rate and batch size and selected the pair with the highest score on the validation set.

4.3 Evaluation Metrics

In selecting the proper metrics to evaluate the results of potential models on train and test data, we recognized that accuracy might potentially be problematic because of its oversimplification and its particular vulnerability in dealing with imbalanced data. In our case, however, because we have greatly mitigated the problem of class imbalances by selecting the optimal threshold in transformation of the response variable, we continued to consider accuracy a valid metric for model evaluation. In addition, we investigated common and proper metrics used by medical

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

journals especially in cases to predict mortality to expand the number of metrics to be used. We decided to include F1 score and AUC due to their ubiquity as metrics in scientific literature [3,4]. After choosing the best model, we further used SHAP to interpret the feature importance. SHAP value computes the marginal contribution of each feature from the baseline prediction (average) to its actual prediction. SHAP value can also be used to explain the global importance by simply averaging the SHAP values across each example; features with large mean absolute SHAP values are considered important.

5. Results and Analysis

5.1 Advanced Feature Engineering and Selection

5.1.1 Feature Selection

Using Lasso regularization, we removed 4 predictors whose coefficients shrunk to 0 - eosinophiles, monocytes, race2, as well as male (Figure 15). The feature importance scores from Random Forest were shown in Figure 16. Predictors with scores lower than the 0.007 threshold were removed. In total, 20 features were left in the subset.

5.1.2 PCA Transformation

Based on the proportion of total variance with different number of PCA, we chose the first five principal components for the original set of predictors which could explain 90% of total variance (Figure 17) and the first 7 principal components for the subsetted predictors which could explain 89% of the total variance (Figure 18).

5.2 Modeling

5.2.1 Model comparison in the original dataset

Based on the original data set, we ran all the models above, evaluated every model using the same metrics: accuracy, F1 score and AUC (Figure 22), and drew the ROC plot. We could see that our baseline model (KNN) achieved the lowest accuracy and F1 score on test data. For our problem, we focused on F1 score and AUC metrics, which offered a better evaluation of model's overall performances, and we could see that among all the models, random forest and neural networks seemed to achieve best performance on test data.

	Train Accuracy	Train F1 Score	Train AUC
knn	0.797210	0.810610	0.872098
logit	0.803687	0.816017	0.871288
tree	0.812407	0.828121	0.876050
rf	0.999751	0.999772	1.000000
adaboost	1.000000	1.000000	1.000000
svm	0.823244	0.835897	0.891874
nn	0.792476	0.807843	0.858495

	Test Accuracy	Test F1 Score	Test AUC
knn	0.777390	0.793724	0.839838
logit	0.796813	0.810056	0.859104
tree	0.781873	0.799817	0.829420
rf	0.795319	0.813097	0.861251
adaboost	0.787351	0.803135	0.848719
svm	0.791335	0.807533	0.849274
nn	0.793825	0.816976	0.859249

Table 9 & 10. Summary of all model performances on original training dataset (9) and testing dataset (10).

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

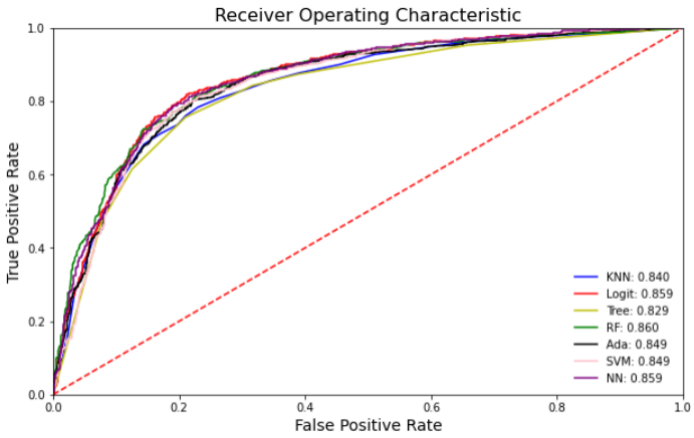

Fig 22. ROC curves of model performances on original dataset

5.2.2 Model comparison between datasets

Random Forest and Neural Network, which were the two best models on the original data, were utilized on the other three different datasets, and their performances were displayed in the table below.

We found that the Random Forest model on the original dataset was the only combination that had F1 score higher than 0.8 and AUC score higher than 0.86. Therefore, to achieve the best prediction performance, medical researchers should use the full dataset, which captured more information, and a random forest model may be a good choice.

However, all the other combinations also achieved very similar accuracy, F1 score and AUC scores. We noticed that the subset dataset achieved similar performances, while reducing more than half of the features from the original dataset (42->20). Therefore, in real practice, clinicians could use the subset data to do prediction, which would largely reduce the work and difficulty associated with data collection.

Principal component analysis largely reduced the data dimension and computational complexity, while achieving similar performances. The PCA after feature selection data achieved similar performance in both random forest and neural network models as a feature selection method, while it reduced the number of features from 20 to 7. The PCA after the original data performs slightly worse than the original data, but the number of features is reduced from 42 to 5. Therefore, when the amount of data is huge, PCA could largely reduce the computational time and fasten the process.

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

	Test_Accuracy	Test_F1 Score	Test_AUC
rf_full	0.794323	0.812187	0.860664
nn_full	0.796813	0.815718	0.859292
rf_sub	0.786355	0.805618	0.850724
nn_sub	0.781873	0.798898	0.851562
rf_pca	0.779382	0.797624	0.837283
nn_pca	0.780378	0.798538	0.844219
rf_pca_sub	0.785857	0.804545	0.845493
nn_pca_sub	0.782371	0.799265	0.848016

Table 17. Model performances comparison of different datasets

5.2.3 Feature importance analysis on the best performed model and dataset

Feature Importance from Random Forest:

From the analysis above, we found that the random forest model on the original dataset performed the best. Based on this model, we drew a feature importance plot from this random forest model. From this plot, we could see that age was the most dominant feature in the random forest model. Many other features, such as systolic_blood_pressure, pulse_pressure and etc., also had important impacts on our response.

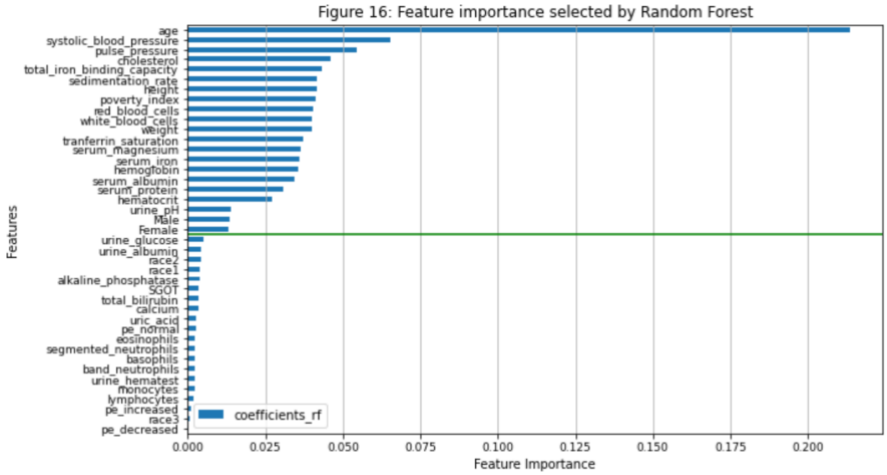

Fig 16. Feature Importance from Random Forest

SHAP Values:

Since traditional feature importance only tells us which features are the most important on a global level, the SHAP values can explain the magnitude and direction of each feature on every single data point. SHAP values relax the linear model assumption of LIME methods and can examine the interaction among features. Based on the summary plot, we observed that age, gender, systolic blood and pulse pressure had the most significant impact on mortality. In particular, large age, high blood pressure and pulse pressure and male would have a high likelihood to die in the next 19 years, which makes sense. If someone is quite old, he or she is less likely to survive the next 19 years. Blood pressure and pulse pressure are two elementary measurements of people's health, which may directly reflect people's likelihood of living or

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

death. The difference between genders matches some previous research that women tend to live longer than men when we hold other conditions constant.

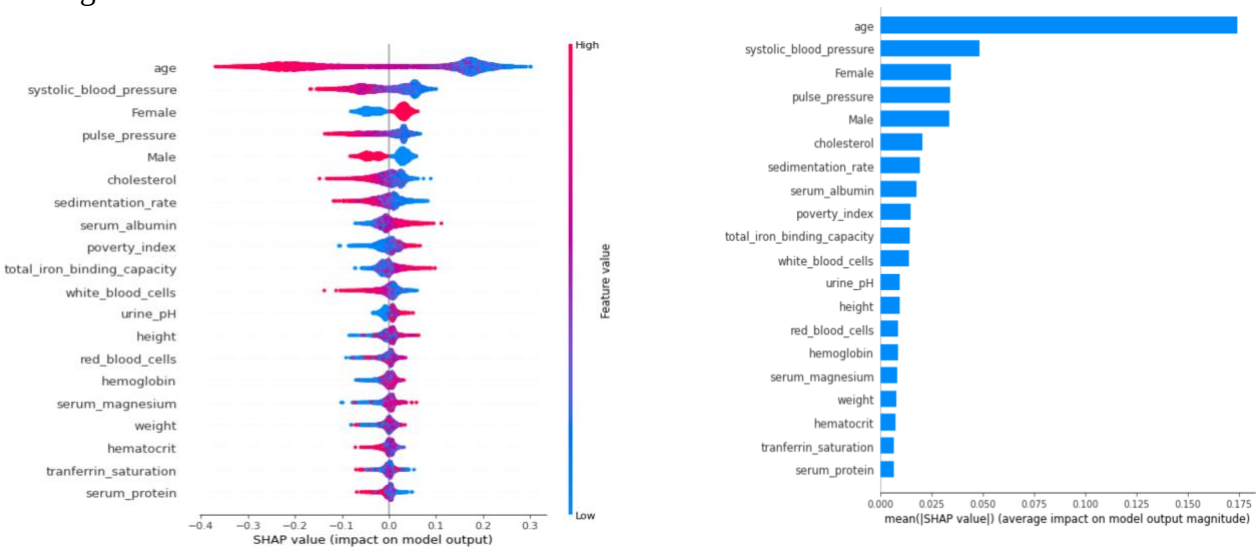

Fig 29 & 30. Summary plots of SHAP value

Interaction Plots:

In order to select the most important four features from the SHAP value plot, we constructed the interaction plots among them. Positive SHAP values showed contribution to predicting to live at least 19 years. Negative SHAP values showed contribution to predicting the death within 19 years.

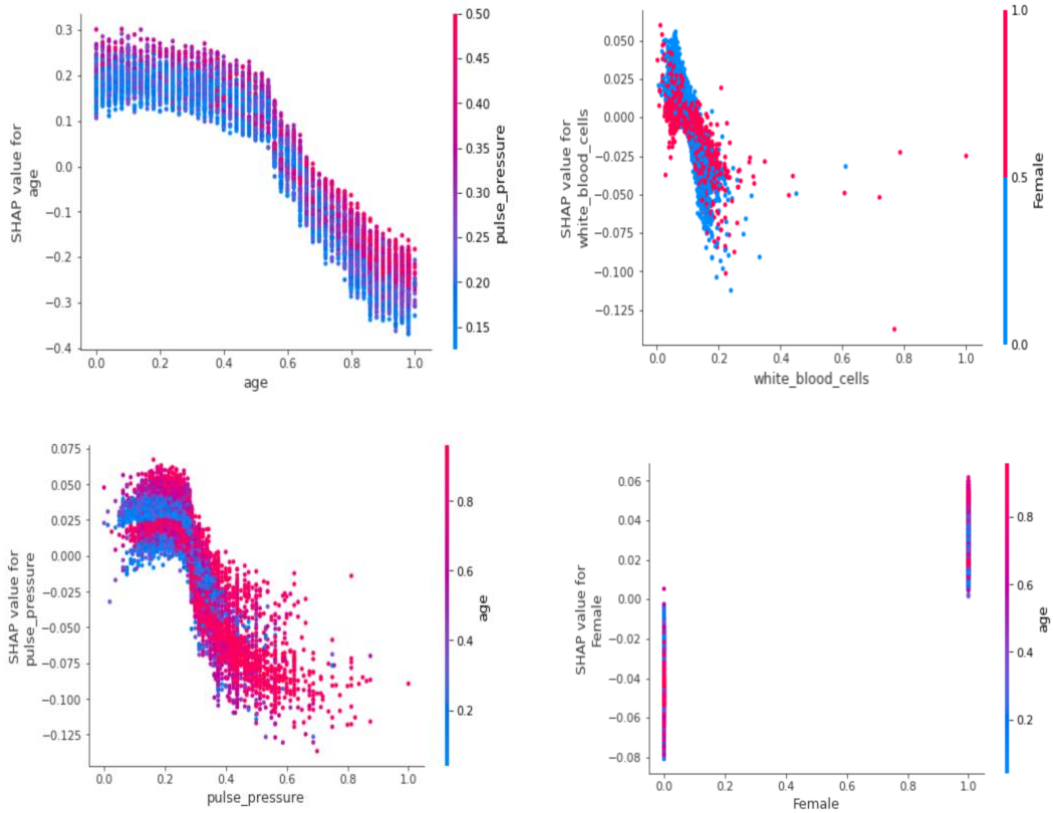

Fig 31-34. Dependence plots of SHAP value for age, white blood cells, pulse pressure and female

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

From these plots, we could find some interesting patterns. As age increases, its contribution to predicting the death within 19 years increases, which makes sense. As pulse pressure increases, its contribution to predicting the death within 19 years increases as well. When whilte blood cells levels were small, the contribution of male's white blood cell to the probability that he lives at least 19 years is bigger than that of female's. However, when whilte blood cell levels were higher, the contribution of male's white blood cells to predicting death within 19 years was greater than that of female's. Compared to makes, females generally had higher contributions to the prediction that a person can live more than 19 years. As age increases, whether the person was female or not became more important in predicting whether he would die within 19 years. As age increased, females seemed to have a higher possibility of living more than 19 years, while males seemed to have a higher probability of death within 19 years.

6. Discussions and Conclusions

In conclusion, a random forest model based on the original data set (after data pre-processing) would best predict a person's risk of dying. Among all the predictors, age, whilte blood cells, pulse pressure and gender seemed to be the most important four features with some interaction effects between them. However, recognizing that a goal of this project was to create a prediction tool of mortality that would be useful to clinicians as well as medical researchers and personnel, we sought to limit the number of predictors used in the model and include only those most relevant and significant. We did this with the understanding that in medical research and clinical settings it is often impractical or prohibitively expensive to collect too many categories of features from patients. Limiting the number of predictors including clinical measures (e.g. urine PH) used in the model would also be helpful to ensure the consistency in data because by reducing burden in data collection (for example through reducing the number of visits required of participants for the study) we could reduce dropout rate of participants, and reduce errors in data due to complicated study administration. Thus, in reality, for the convenience of clinicians to collect data, we could use a subset of 20 features as described above to perform the prediction analysis while achieving similar performance. PCA was also a good method in practice for a large dataset as it greatly reduced data dimension and computational time while achieving similar results.

However, there are limitations regarding our approach, analysis, and/or data used. To begin with, due to significant missingness in many of our predictors, we lost a significant amount of information as we were forced to eliminate some predictors from our analysis or convert them to variables signaling only their missingness. In addition, because we converted our response variable from continuous to categorical, we also could have lost information by using a categorical variable for evaluation of mortality risk; after all, two participants dying 3 years and 15 years since the first examination of the study were categorized the same in our approach although it would be useful for us to learn the reasons explaining this difference. Consequently, future studies could seek to institute better experiment administration to mitigate issues regarding data missing. Since the data we used represented only levels of predictors, it could also be helpful to study the levels of these variables at different followups to understand how the changes of predictors influence mortality risk. Finally, while our dataset did not have a particular focus on patients with a specific type of disease, it would be interesting to validate or test our models in a disease specific context.

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.

7. References

1. Mayo Foundation for Medical Education and Research. (2020, December 22). Complete blood count (CBC). Mayo Clinic. Retrieved December 1, 2021, from https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919.
2. Zhang, S., et al., Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 2017. 8(3): p. 1-19.
3. Yan, L., Zhang, HT., Goncalves, J. et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2, 283–288 (2020).
4. Subudhi, S., Verma, A., Patel, A. B., Hardin, C. C., Khandekar, M. J., Lee, H., McEvoy, D., Stylianopoulos, T., Munn, L. L., Dutta, S., & Jain, R. K. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. Npj Digital Medicine, 4(1).

*Note: All figures are in the Colab. We only chose some figures to be shown in the report.