南開大學
Nankai University

Statistical Computing

Group work

Echocardiogram Prediction

Reporting time：

Group 7 14-Dec-18

Members：

Yin Haoyu	2120180128
Wang Yinge	2120180125
Su Minmin	2120180118
Wang Xiaoyuan	2120180123

Abstract

In order to find out the factors that affect survivors over the past year, and to fit and predict the data.So first, we did the data visualization and pre-processing work. Intuitively observing the individual's survival distribution and finding out factors which have impacts on survival. In the second and third parts, we use logistic regression, decision tree model, random forest and neural network to fit the data after multiple imputation, The ROC curve and the confusion matrix and cost sensitive are used to judge its validity.However, the model exist "over-fitting" because of synthetic samples generated by SMOTE.Thus, in the fifth part, the data is analyzed again for survival.we find three factors that have great influence on survival time.Finally, the four models of data fitting are compared to find that the random forest is the optimal model.And through survival analysis, we found out factors which have the greatest influence on survival time.

Keywords:data visualization,logistic regression,decision trees,random forest,neural network ,survival analysis

Contents

1 Introduction

All the patients suffered heart attacks at some point in the past. Some are still alive and some are not. The survival and still-alive variables, when taken together, indicate whether a patient survived following the heart attack.

The main methods of disease prediction can be divided into three categories: classical regression method, machine learning method and deep learning method.

Traditional disease risk predictions are based on the Cox proportional hazard regression model and the logistic regression model.

Feature selection in the field of machine learning and supervised learning modeling methods are increasingly used for disease prediction problems. Some machine learning methods can improve the interpretability of predictive models, such as decision tree methods.

Our paper not only combines classical regression and machine learning methods for prediction, but also innovatively uses survival analysis, which considers both the results and the length of time each observation has.

Our ultimate goal is to find out the factors that affect the survival of patients with heart disease, and build models to predict the survival of heart disease patients. There are several methods to achieve our goal, logistic model, random forest, neural network and survival analysis. Before analyzing the data, we did the data visualization and pre-processing work first. Intuitively observing the individual's survival distribution and finding out factors which have impacts on survival. After the data was complemented by multiple interpolation methods, we built different models for prediction.

Our data analysis work first introduced theoretical basis of the logistic model and the decision tree model and then fitted three models for analysis. The ROC curve and the confusion matrix are used to judge models' validity. The AUC value of our model is nearly 1. In addition, our model has the correct prediction for all the response variables of zero.

We also use random forest and neural network to analyze our data. AUC curve is plotted to measure the model effect. Compared with the previously established models, these two models have some improvements. However, the model exists "over-fitting" problem because synthetic samples generated by SMOTE.

Our last and also the most important part is survival analysis, we find three factors that have great influences on survival time, they are "age-at-heart-attack", "E-point septal separation", and "wall-motion-index", among which "wall-motion-index" has the greatest impact. Therefore, when estimating survival time of patients, we can focus on these three factors.

Through establishing four models and comparing their advantages and disadvantages, we found out that the best model is random forest to predict patients' survival for this problem. And through survival analysis, we found out factors which have the greatest influence on survival time.

2 Data isualization and data preprocessing

Before establishing models and analyzing the data, we first did some data preprocessing and visualization work on the original data, and visually observed whether the original data is missing, wrong or not available. Using functions in the "ggplot" package to initially observed whether the factors we care about have impacts on the survival of heart attack patients visually.

Over the next, we created complete data sets by multiple imputation to generate possible values for missing values. The analysis process corresponding to the multiple imputation data set generates an output for data set and generates a converged output containing the result estimates when the original data set has no missing values. These aggregation results are usually more accurate than those provided by a single interpolation method.

After deleting the erroneous data and manually replenishing the data by means of multiple interpolation, the interpolation data is used to build models for analysis and prediction.

2.1 Data Visualization

2.1.1 Visualizing the data structure of the original data.

First, we imported the raw data into the R software. We can see the name and type of thirteen variables visually by drawing a structure diagram of the original data.

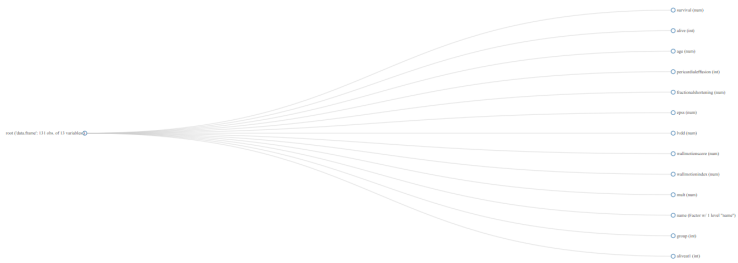

Figure 1: A structure diagram of the original data

2.1.2 Visualizing the missing data value

We can draw a conclusion from the missing value plot that almost all of the thirteen response variables have missing values, and some of the response variables have the number of rows with missing values is up to 44.27%.

Therefore, before analyzing the data, we used the method of Multiple Imputation to artificially fill the missing data.

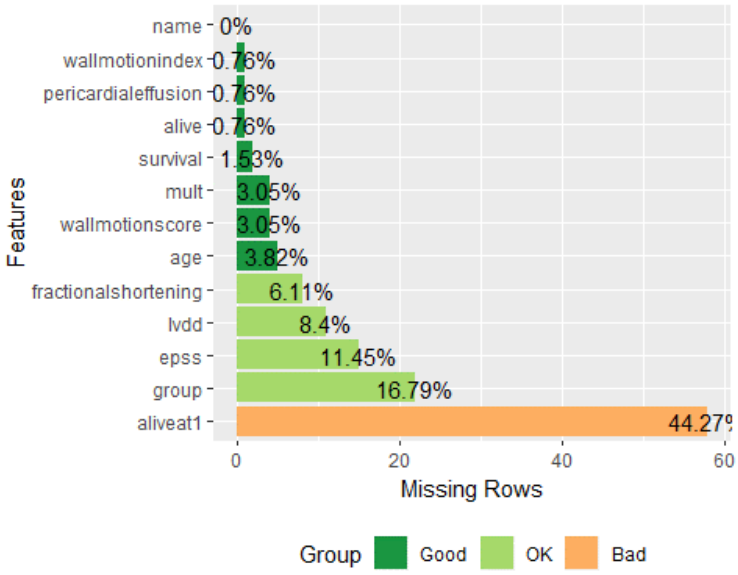

Figure 2: The missing value plot

2.1.3 Visually analyzing the original data

Red dots indicate individuals who are died, while blue dots indicate individuals who are still surviving. Different sizes indicate different age groups, and the life expectancy of individuals with different ages can be clearly observed from the pictures.

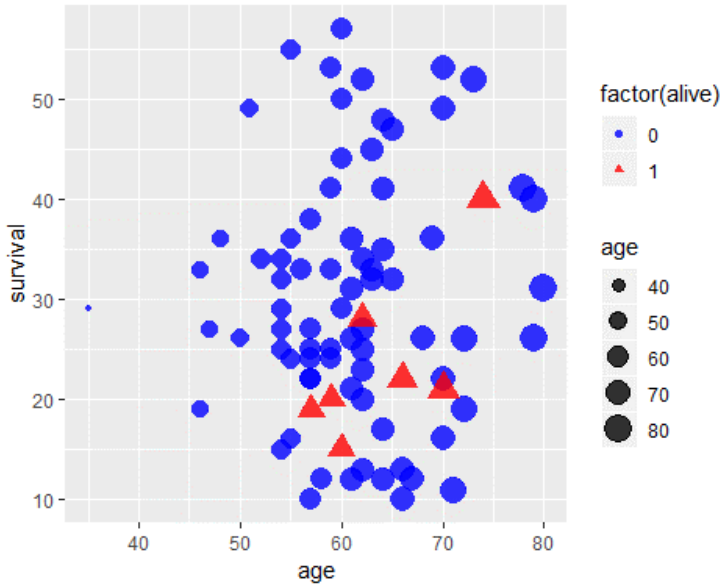

Figure 3: Age-Survival

The following pictures depict the density function of "epss", "fractional-shortening", "lvdd", "pericardial-effusion" and "wall-motion-index" these five variables. The red shaded part represents the distribution of

dead individuals while the green shaded part represents the distribution of living individuals. It can be seen from the graph that the samples with different living conditions have different kurtosis and skewness of different affecting factors, which can preliminarily reveal that these different factors have impacts on whether individuals are still surviving or not.

Figure 4: Density curve of different variables

As the first picture of **Figure** 4 plots the distribution of the variable E-point separation (a measure of contractility), it can be seen that for dead individuals, the peak value is around 9, while for surviving individuals, the peak value is around 20.

2.2 Data Preprocessing

2.2.1 Eliminating and Renaming Variables

We use "wall-motion-index" instead of "wall-motion-score". "Mult", "name", "group" are meaningless, so we ignored them. "Alive-at-1" isn't the factor we care about, so eliminated it. And then there are seven independent variables, while "alive" is the only dependent variable. For convenience, we use "wmi" ,"pe" and "fs" instead of "wall-motion-index","pericardial-effusion" and "fractionalshortening".

2.2.2 Multiple Imputation

We removed the insignificant predictive variables from the original data, and used functions in the "mice" package to do multiple imputation and then artificially fill in missing values.

After the data of surviving individuals with a survival period of less than 12 months was excluded, we read the interpolation data.

2.2.3 Visually analyzing the interpolation data

Visually analyzing the data whose insignificantly predictive variables, missing values, and erroneous data were eliminated.

You can see the distribution of dead and live individuals with different ages from the graph.

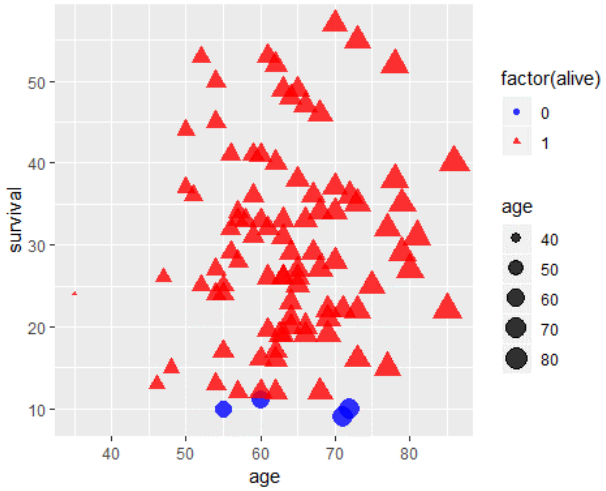

Figure 5: The distribution of dead and live individuals with different ages

The distribution of these five variables indicate that these variables have impacts on individuals survival.

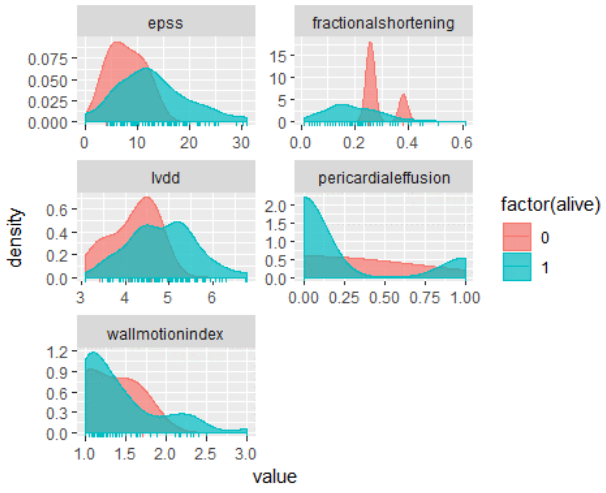

Figure 6: The distribution of five variables

3 Data analysis

3.1 Logistic Regression

3.1.1 Theoretical basis

Assuming that the response variable Y is a two-class variable and there are K factors affecting Y, then model:

$$ln(\frac{p}{1-p}) = g(x_1, ..., x_k) \tag{1}$$

The above formula is a logistic regression model of binary data. Among them, k factors are called covariates of logistic regression models.

3.1.2 Model fitting

The initial model is model 1, The model obtained by stepwise regression of the selected variable is model2, and the AIC value of Model 2 is reduced by about 1.7 relative to Model 1. and there is no difference in the degree of fitting of the two models from the variance analysis of the two models, so that the model 2 is considered to be better than the model 1.

$$model1 : log(\frac{p}{1-p}) = \beta_0 + \beta_1 age + \beta_2 pe + \beta_3 fs + \beta_4 epss + \beta_5 lvdd + \beta_6 wmi \tag{2}$$

Which$p = p(alive > 1year|age, pe, fs, epss, lvdd, wmi)$

$$model2 : log(\frac{p}{1-p}) = \lambda_0 + \lambda_1 pe + \lambda_2 fs + \lambda_3 epss + \lambda_4 lvdd + \lambda_5 wmi \tag{3}$$

The regression diagnosis of the model 2 is as follows: the variable pe in the model 2 is not significant ,and the influence diagram of the **Figure** 7 shows that the model 2 has high influence points and high leverage points.

$$model3 : log(\frac{p}{1-p}) = T_0 + T_1 age + T_2 fs + T_3 epss + T_4 lvdd + T_5 wmi \tag{4}$$

Since model 2 has high influence points ,We perform robust logistic regression on the data to get model 3. all variables test is significant. From the ROC curve, the prediction accuracy of Model 3 is 0.899.

3.2 Decision Trees

3.2.1 Theoretical Basis

The classification decision tree is a tree structure that describes the classification of the instance Venus. The decision tree consists of nodes and directed edges. There are two types of nodes: internal nodes and leaf nodes. An internal node represents a feature or attribute, and a leaf node represents a class.The decision tree is used to classify, starting from the root node, testing a certain feature of the instance, and assigning the

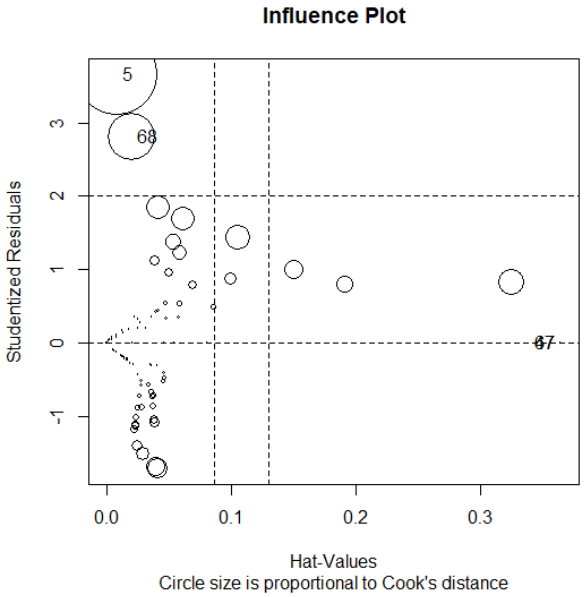

Figure 7: Impact diagnosis chart

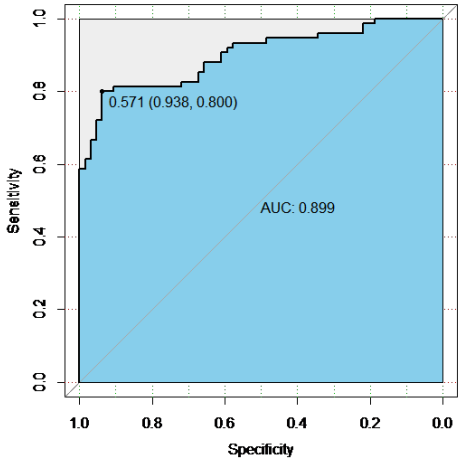

Figure 8: ROC curve

instance to its child nodes according to the test result; at this time, each child node corresponds to a value of the feature. The instance is tested and assigned recursively until the leaf node is reached, and finally the instance is assigned to the class of the leaf node.The construct is shown as **Figure** 9.

3.2.2 Model Fitting

The data is divided into a training set and a test set, and the ratio of the division is close to 3:1. The model 4 is obtained by fitting the decision tree to the training set. From the tree structure of Model 4: its variables are epss, lvdd, wall motion index. Then, from the confusion matrix of model 4 : For the data with a response variable equal to 0, all fit correctly.

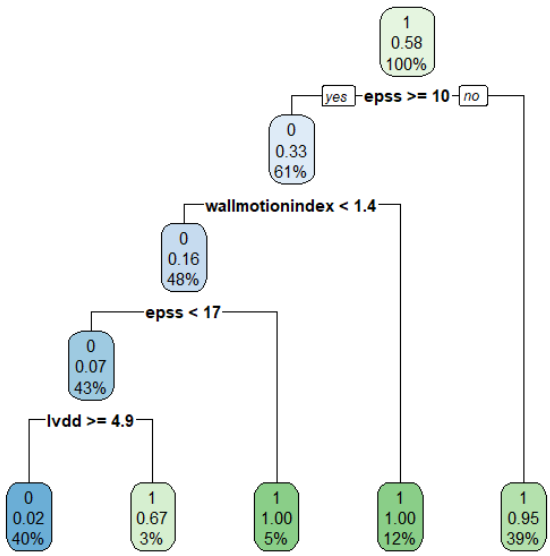

Figure 9: Decision tree fitting model

Table 1: The confusion matrix

Referencece/Prediction	0	1
0	20	0
1	1	14

Finally, the AUC = 0.967 is obtained from the ROC curve on the left. Therefore, the prediction accuracy of Model 4 is not bad.

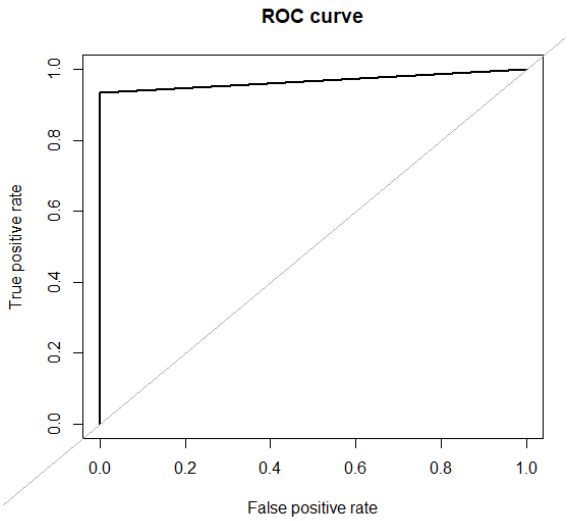

Figure 10: Decision trees ROC crve

4 Other models

In this part, our report will introduce some other models.

4.1 Random Forest

First,we use random forest. The model framework is shown as **Figure** 11. We construct 500 trees, for each CART subtree,we choose the sample and features randomly to train it.From these trees,we can finally get the results by voting.

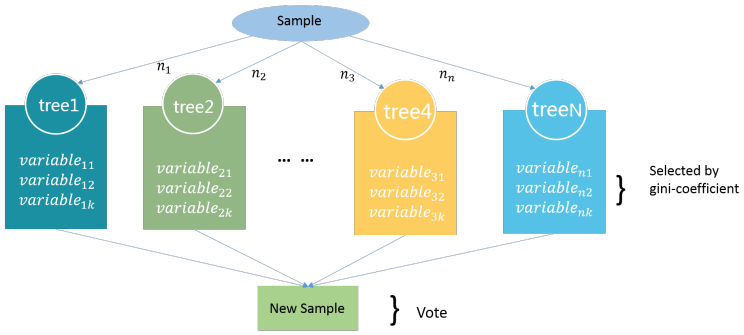

Figure 11: The framework of random forest

The results of random forest is shown as **Figure** 12. AUC reaches to 1 and ROC curve performs perfectly.

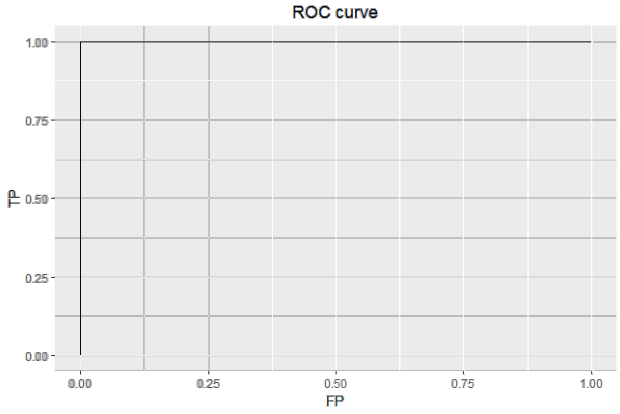

Figure 12: The result of random forest

Here we need explain why the ROC curve is not smmoth. Excessively small testing sample size and extremely high prediction accuracy will both make the curve jump. We use nearly a third of the total samples as the testing samples, so our testing sample size is enough. However, the prediction accuracy is unexpectedly high, so the curve is not smooth.

The forest also tell us some infomation about he variables.We use gini-coefficient to reflect the importance

of the variables. The samller the score, the more important the variable is. Gini-cofficient is defined as:

$$Gini(D) = 1 - \sum_{k=1}^{y} p_k^2 \tag{5}$$

Which y means the number of class and p_k means the the probability that the sample belongs to this class.

The result is shown as **Table** 2.From the table we know that it is necessary to give priority to variables such as wall-motion-index and age.

Table 2: The importance of variables

Variable	Gini-coefficient
pericardial-effusion	2.02
wall-motion-index	4.58
age	5.53
fractional-shortening	10.05
epss	16.96
lvdd	17.42

4.2 Neural Network

The construct of the neural network is shown as **Figure** 13.

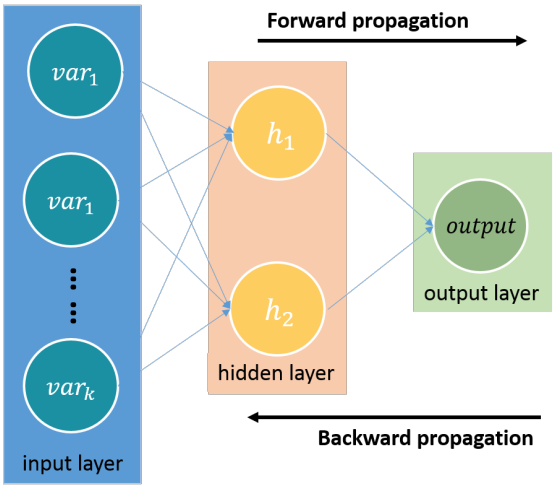

Figure 13: The framework of neural network

We use cross entropy as the loss function and forward propagation and backword propagation to calculate the result and estimate the parameters. The network only has one hidden layer and five hidden nodes.

The results are shown as **Figure** 14. AUC reaches to 0.93 and ROC curve also performs perfectly(a bit weaker than random forest).

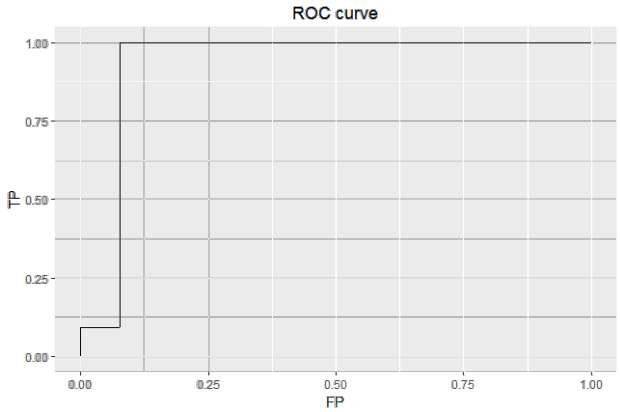

Figure 14: The framework of neural network

4.3 Disscussion

If we only use ROC curve to measure the performance of the model,there are some weakness.If a person who can not live more than one year is predicted as a person who can live more than one year,we think the case is very serious.Contrarily,we think the case is acceptable.So we define the different cost for each mistake,the confusion matrix is shown as **Figure** 15.

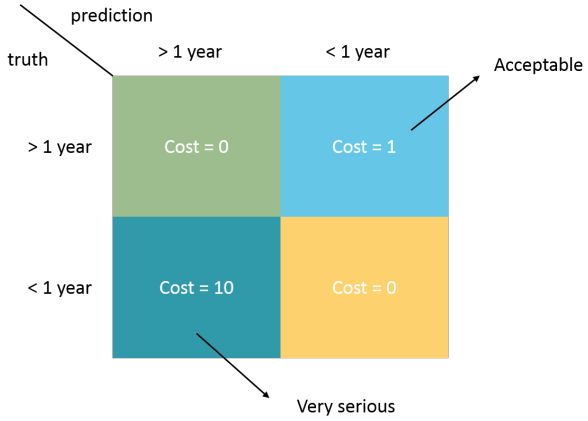

Figure 15: The confusion matrix

Then,we define cost sensitive:

$$L = \frac{n_{10}cost_{10} + n_{01}cost_{01}}{n} \tag{6}$$

The cost sensitive of random forest is 0.166 and the cost sensitive of neural network is 0.93.It also seems perfect.Our model seldom makes mistake and perform perfect like the textbook.

In fact,we only have 4 positive samples,we use smote to generate other 60 samples.We can see that the generated samples are very similar.So our model may just remember these samples instead of finding the rules.Whem it receive an actual new sample,the model may not know how to deal with it.

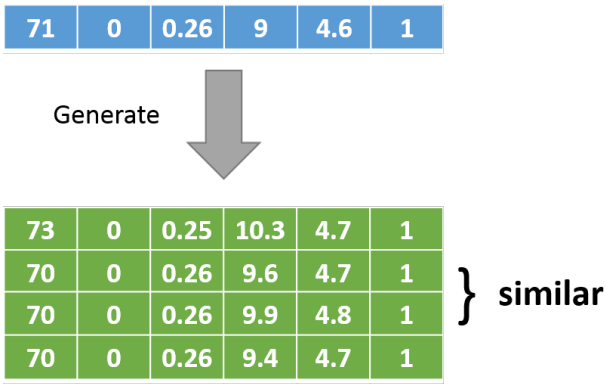

Figure 16: Generayed samples

So we should find other method to help us diagnosis,the report will introduce survival analysis in the next part.

5 Survival analysis

5.1 Applicability

The survival time of the patients we observed can be divided into the cases as followed. Because the time we observed is limited, so only part of the patient's survival trajectory can be completely covered, and for other samples that are not fully observed, it is called censored data. In this study, we observed the lifetime and survival status of 132 patients at the cut-off time.

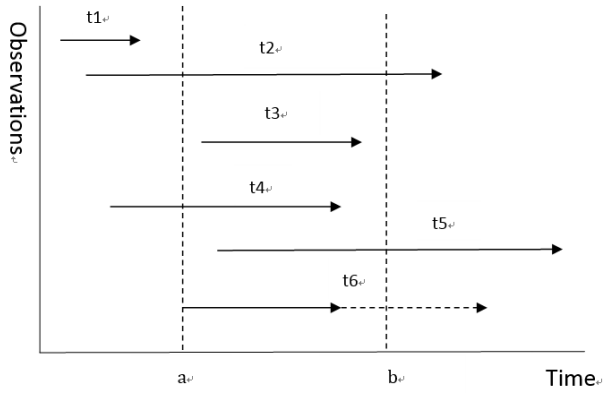

Figure 17: Observation of patients' survival time

Since the patients' survival time is censored data, the direct use of models such as regression equations can cause large errors in the prediction. Therefore, in the data processing, patients are divided into two groups according to whether they could live for a year or not. This way, we chose two classification models, which were introduced before. But those only used part of original data information, so we also used survival analysis to make full use of the original data. Since there are many samples in the data with default values, in

order to ensure the good results, we only select the complete samples for analysis.

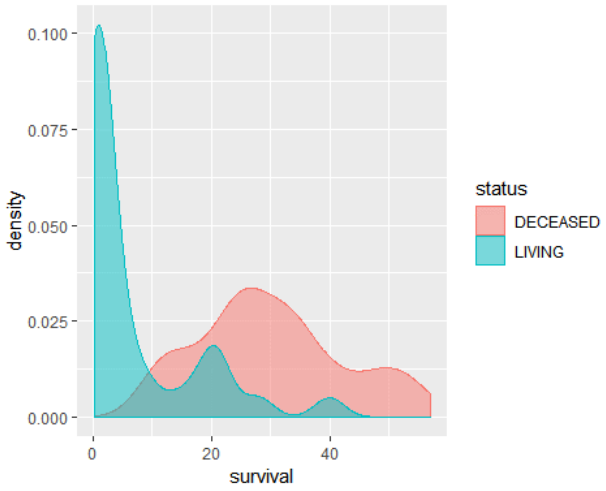

Figure 18: The nuclear density curve of survival time

We first plot the nuclear density curve of the survival time of patients with different survival status. The survival status is divided into deceased and living. It can be seen that the survival time of the deceased patients is symmetrical, and the survival time of the living patients is right-biased. It means that the patients who have been treated can be guaranteed a longer survival time before they die.

5.2 Simple KM survival analysis

The **Figure** 19 shows the estimated value and confidence interval of the survival curve of the deceased patients. The curve shows a blue solid line, and the 95% confidence limit shows a grey background. It is easy to see that the survival rate of the deceased patients tends to zero over time.

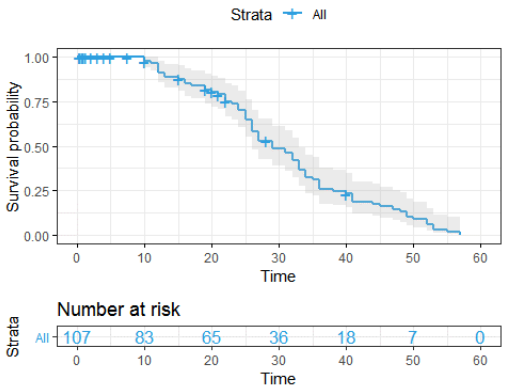

Figure 19: Survival curve

Some researchers tend to generate cumulative mortality curves rather than survival curves, which show cumulative probability of experiencing events of interest. From **Figure** 20 we can estimate the possibility of death at a certain point in time.

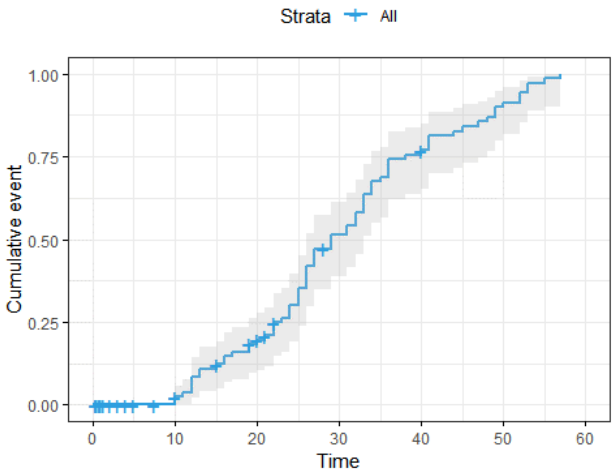

Figure 20: Cumulative mortality curve

5.3 Classified KM survival analysis

First, we can look for variables that differ greatly from patients in different survival status, such as age. According to the boxplot, we can see that the mean values of the deceased and the living are quite different. Because the p value is less than 0.05, it indicates that the age of patients in different survival status has different expressions.

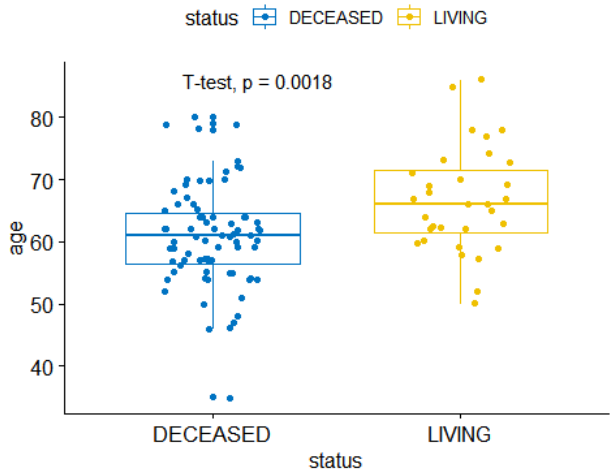

Figure 21: The boxplot of the deceased and the living

It can be determined by the function that the best segmentation point of age is 65 years old, and the survival curve is drawn by dividing 65 years old. With the extension of time, the difference in the rate of decline in the survival rate of patients in different age groups is still relatively large.

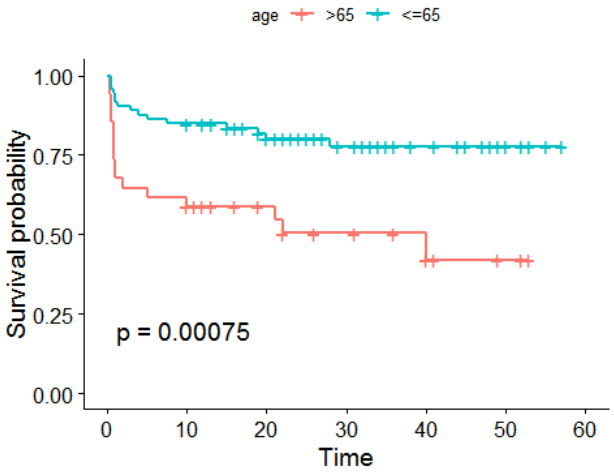

Figure 22: Survival curve in different groups

5.4 Cox survival analysis

The cox survival analysis plays a very important role in the survival analysis. If we do not know the type of distribution of survival time, but also want to analyze the impact of multiple risk factors on survival time, we can use the Cox model. It is able to study the relationship between predictors (covariates) and "time-events" through risk functions. The Cox model can overcome the current different living conditions of the patients and combine them with survival time for analysis. The basic form of the Cox model is followed:

$$h(t|x) = h_0(t)exp(\beta_1 x_1 + ... + \beta_p x_p) \tag{7}$$

Through the stepwise Cox regression we finally get the following results.

Table 3: The result of stepwise Cox regression

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age	0.8139	1.80479	0.02154	3.778	0.00158
epss	0.04566	1.04672	0.02165	2.109	0.034907
wmi	1.18291	3.26384	0.39005	3.003	0.002424

As can be seen from the results, the coefficient of each variable is significant. For each additional unit of age, the patients' mortality rate will increase by 0.085. For each additional unit of epss, the patients' mortality rate will increase by 0.047. For each additional unit added by wmi, the patients' mortality rate will increase by 2.263.

Thus, the Cox model can find which variables have a greater impact on the survival time of patients, and quantify the extent to which each variable affects the survival time.

6 Conclusions

We have already built four models. We can see that the results of neural network and decision tree are better than logistic regression. In theory, the results of random forest should be the best. However, due to the imbalance of data, we came to the results of over-fitting.

Table 4: The result of different models

Model	LOGISTIC Regression	Neural Network	Decision Tree model	Random Forest
AUC	0.899	0.93	0.967	1.00

Through survival analysis, we find three factors that have great influence on survival time, among which wall-motion-index has the greatest impact. Therefore, when estimating survival time of patients, we can focus on these three factors.

Table 5: The result of survival analysis

Variables	Variances
age-at-heart-attack	0.085
E-point septal separation	0.047
wall-motion-index	2.263