Final Project

University of Minnesota

Yisak Kim, Sihyeon An, Hyeonung Kwang

Dec 2021

Introduction

Heart disease is one of the serious diseases that modern people have. Various factors can cause heart disease. Our research goal is to predict heart disease and analyze which factors can significantly influence the risk of heart disease. There are 9 variables that can influence heart diseases: age, gender, chest pain type, resting blood pressure, cholesterol level, fasting blood pressure, resting electrocardiogram results, maximum heart rate and exercise induced angina.

Method

In the dataset, there are missing values on the 3 predictors: maximum heart rate, chest pain type and heart disease. We used two imputation methods: simple imputation and iterative regression. For the simple imputation, we deleted missing observations on the categorical variables. Then, for the continuous variable, we imputed the missing values with a mean value of the variable. For the iterative regression imputation, we used logistic regression to impute missing values of heart disease; we used multinomial regression to impute missing values of chest pain type; we used linear regression to impute missing values of maximum heart rate. After imputation with both methods, we divided the dataset into the training set and test set. The reason why we divided the dataset is to know whether there is overfitting or not. For the predictive model, we used logistic regression, KNN, and random forest. To improve the performance of logistic regression, we added the penalty with lasso and ridge. We chose the best logistic regression model which has the lowest error rate or highest AUC. The second analysis method is KNN which is to classify observations with the nearest values. KNN is a non-parametric classification method, so this method is more flexible than the parametric method. The last model is a random forest. The advantage of using a random forest is that the variance can be reduced by bootstrap sampling. There are two hyperparameters in the Random Forest which are the *mtry* and *B*. The optimal parameters were applied to have the best result.

Results

For simple imputation, the suitable logistic regression model was lasso logistic regression. To find the optimal tuning parameter, cross-validation was used. The optimal tuning parameter is 0.01321941. Table 1 shows the result of the test error rate. As we can see, the lasso logistic regression had the lowest test error rate. For the overfitting problem, it seemed to be fine because the error rate between the training and test set are similar.

<Logistic Regression/Ridge logistic regression/Lasso logistic regression>

Logistic Regression Error rate		
Model	Training set	Test set
Logistic regression without penalty	0.1636364	0.2116402
Ridge logistic regression	0.1568182	0.2222222
Lasso logistic regression	0.1568182	0.2116402

Table 1

We made the roc curve plot to estimate the better fit model among models. Figure 1 shows the roc curve of logistic regression models.

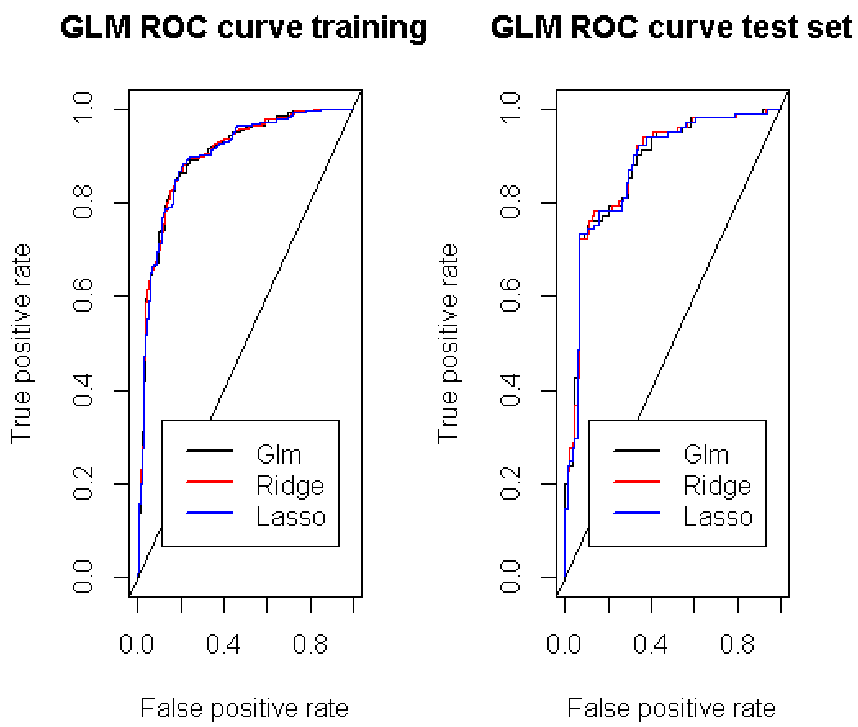

Figure 1

However, it is not clear to figure out the best suitable logistic regression model with the plot. Therefore, we computed the AUC value based on the roc curve. Table 2 shows the AUC values of each model. When we compared the models using AUC value, the ridge logistic regression had the highest AUC value.

Logistic Regression AUC		
Model	Training set	Test set
Logistic regression without penalty	0.9001359	0.8756751
Ridge logistic regression	0.9003269	0.8778128
Lasso logistic regression	0.8985118	0.8751125

Table 2

<KNN>

For the K-nearest neighbor method, we estimated the error rate with a 10-fold cross-validation and validation-set approach. Table 3 and 4 indicates the error rate of the different number of k-nearest neighbor models. Based on the 10-fold cross-validation, when k= 21, it shows the lowest error rate.

KNN 10-fold Cross-Validation Approach Error rate		
K-nearest	Training set	Test set
3	0.1878981	0.3015873
5	0.2356688	0.3015873
10	0.2463312	0.3492063
21	0.2770701	0.2380952

Table 3

Based on the validation set approach, when k=10, it shows the lowest error rate.

Validation Set Approach Error rate		
K-nearest	Training set	Validation Set
3	0.1863636	0.3703704
5	0.2250000	0.3597884
10	0.2750000	0.3492063
21	0.2568182	0.3544974

Table 4

The reason why they have different results is that the validation approach is strongly affected by which observations enter in the training and validation sets. Therefore, the validation set approach may lead to biased estimates of test error rate. Thus, we decided to use the error rate of 10-fold cross-validation.

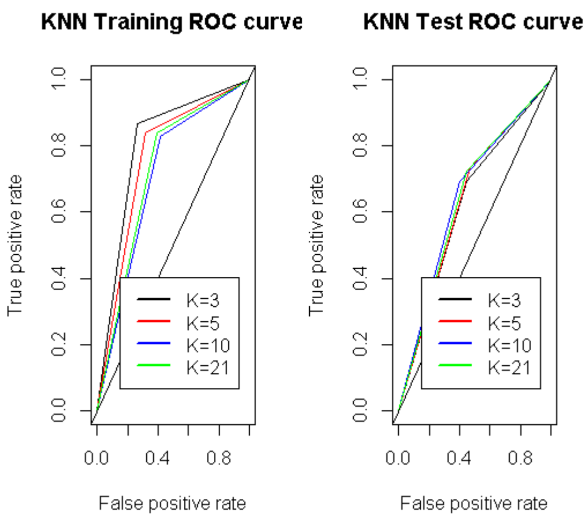

Figure 2 shows the roc curve with different k values. However, it is not clear to figure out the best suitable KNN model with the graph. Therefore, we computed the AUC value based on the roc curve. Table 5 shows the AUC values of each model. As we can see, when k=21, the AUC value is the highest.

Figure 2

KNN AUC		
Model	Training set	Test set
KNN with k=3	0.8031590	0.6249437
KNN with k=5	0.7623132	0.6333821
KNN with k=10	0.7048234	0.6376710
KNN with k=21	0.7242697	0.6397952

Table 5

<Random Forest>

Since we have 9 predictors, we used p/3 as *mtry* value; *mtry* is 3. Table 6 shows the error rate of random forest.

Random Forest Error rate	
Model	Test error rate
Random Forest	0.1957672

Table 6

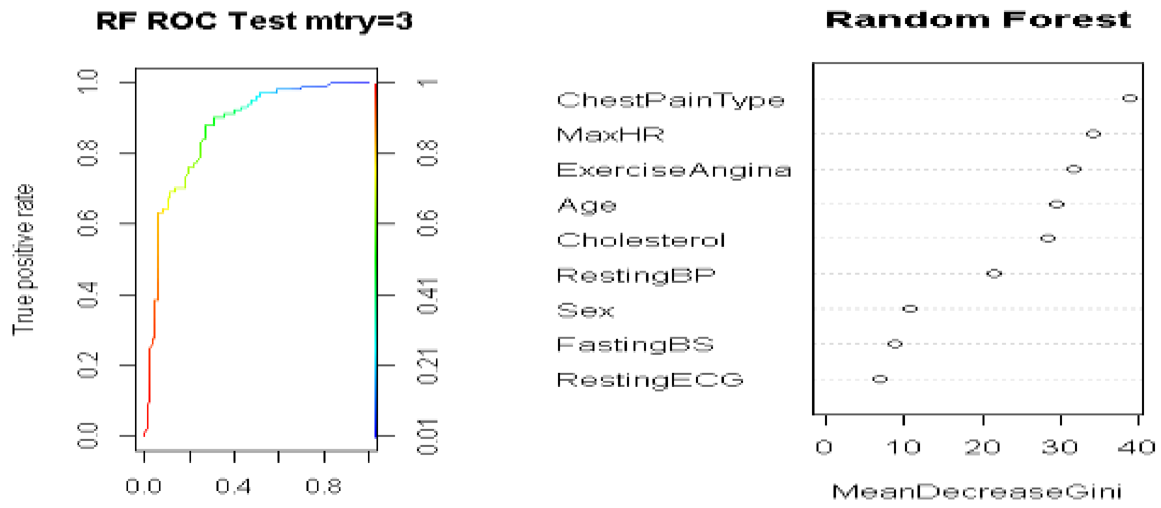

Figure 3

Figure 3 shows the ROC curve of random forest and importance plot. The AUC value of random forest was 0.8690932..

<Comparison 3 models by using error rate>

Table 7 shows the error rate of each model selected by applying the optimal parameters. As a result, the best model using error rate is the random forest model.

Comparison Three models using ER	
Model	Test error rate
Lasso logistic regression	0.2116402
KNN with 10-fold when K=21	0.2380952
Random Forest with mtry=3	0.1957672

Table 7

<Comparison 3 models by using AUC>

Table 8 shows the AUC of each model selected by applying the optimal parameters. As a result, the best model using AUC is ridge logistic regression.

Comparison Three models using AUC	
Model	AUC
Ridge logistic regression	0.9056122
KNN when K=21	0.6397952
Random Forest	0.8690932.

Table 8

<The result of iterative regression imputation>

The same process that we used for simple imputation were applied for the iterative regression imputation. Through the iterative regression imputation, we have a different number of observations and the distributions are shown in figure 4.

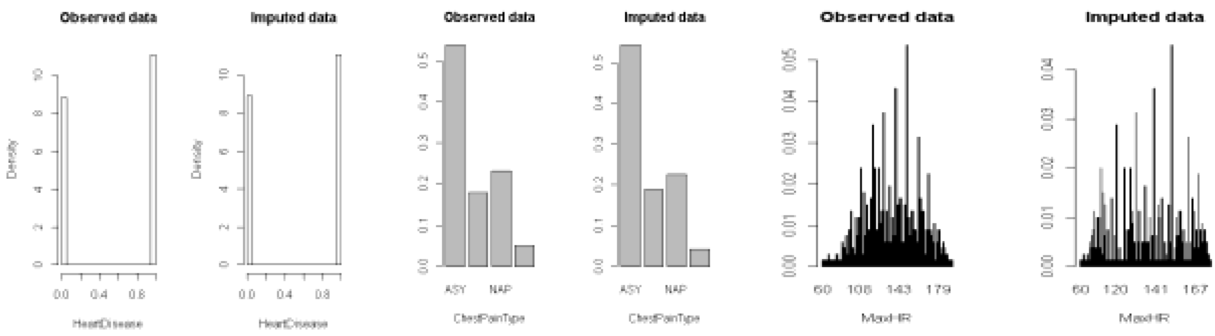

Figure 4

We had 800 observations with Iterative regression imputation while there are 629 observations with simple imputation. Table 9 to 16 and figure 5 to 7 indicate the result of 3 models by iterative regression imputation.

<Logistic Regression/Logistic lasso regression/Logistic ridge regression>

Logistic Regression Error rate		
Model	Training set	Test set
Logistic regression without penalty	0.1500000	0.1791667
Ridge logistic regression	0.1428571	0.1875000
Lasso logistic regression	0.1428571	0.1666667

Table 9

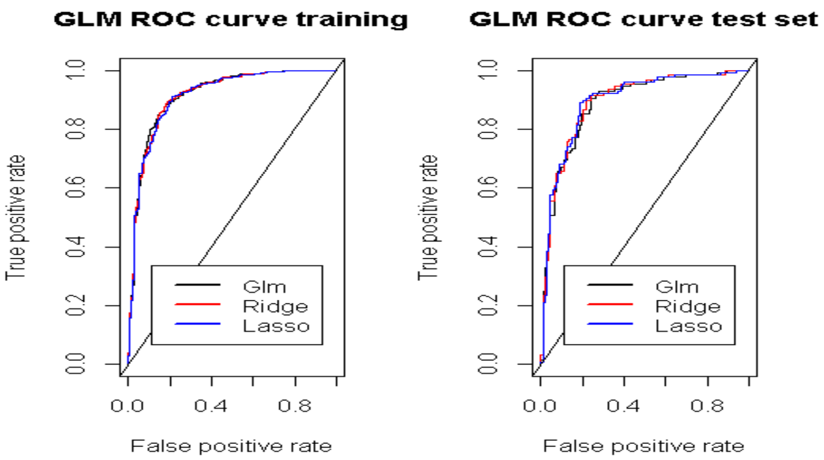

Figure 5

Logistic Regression AUC		
Model	Training set	Test set
Logistic regression without penalty	0.9175529	0.8889509
Ridge logistic regression	0.9164894	0.8943220
Lasso logistic regression	0.9145959	0.8950195

Table 10

KNN 10-fold Cross-Validation Approach Error rate		
K-nearest	Training set	Test set
3	0.1764706	0.3375
5	0.2177722	0.3270
10	0.2490613	0.3250
23	0.2565707	0.3125

Table 11

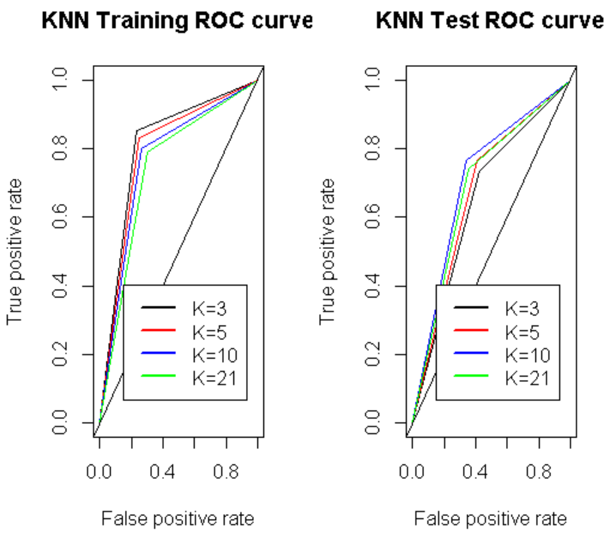

Figure 6

KNN AUC		
K-nearest	Training set	Test Set
3	0.8088296	0.6573661
5	0.7931884	0.6774554
10	0.7671197	0.7131696
23	0.7459795	0.6925523

Table 12

Random Forest error rate	
Model	Test error rate
Random Forest	0.1833333

Table 13

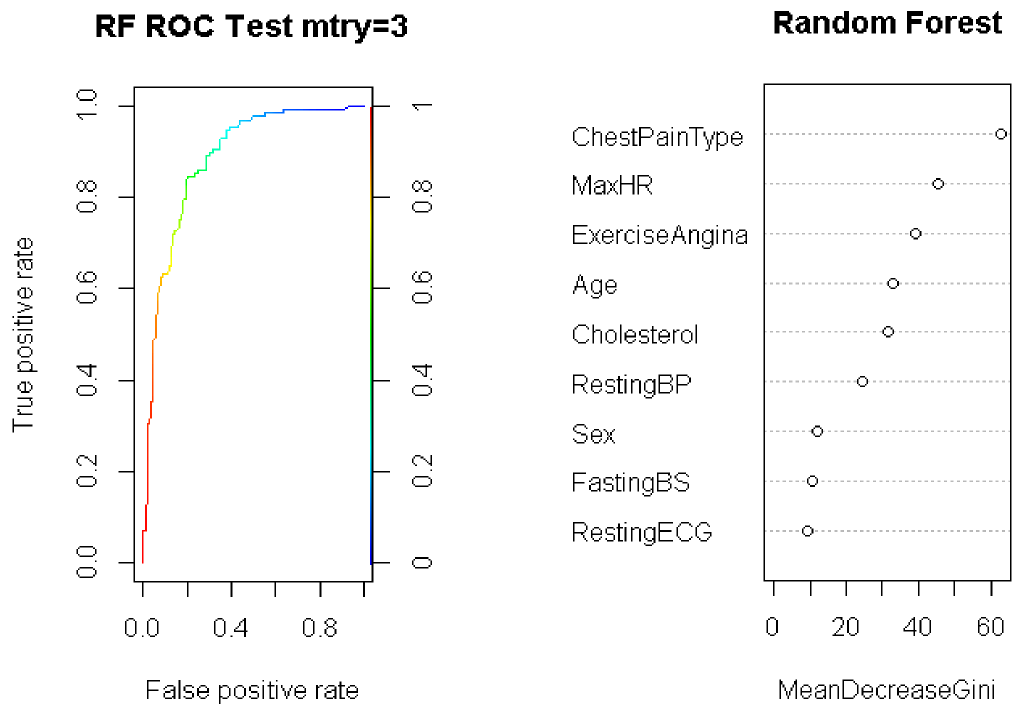

Figure 7

Comparison Three models using ER	
Model	Error rate
Lasso logistic regression	0.1666667
KNN when K=23	0.3125
Random Forest	0.1833333

Table 15

Comparison Three models using AUC	
Model	AUC
Lasso logistic regression	0.8950195
KNN when K=10	0.7131696
Random Forest	0.8844866

Table 16

< Simple Imputation vs Iterative regression Imputation >

Comparison Simple Imputation vs Iterative regression Imputation		
Model	Simple imputation ER	Iterative regression Imputation ER
Lasso logistic regression	0.2116402	0.1666667
KNN	0.2380952(K=10)	0.3125(K=23)
Random Forest	0.1957672	0.1833333

Comparison Simple Imputation vs Iterative regression Imputation		
Model	Simple imputation AUC	Iterative regression Imputation AUC
Lasso logistic regression.	0.9056122	0.8950195
KNN	0.6397952(K=21)	0.7131696(K=10)
Random Forest	0.8690932.	0.8844866

Overall, the best suitable model for predicting heart disease is Lasso logistic regression. Imputation with iterative regression has a lower error rate than simple imputation. This is because simple imputation can cause biased analysis and throw away a lot of information. Also, the sample size is reduced. It consequently causes higher variance and asymptotic results may not hold. Also, there is lower power of test hypotheses.

<Lasso logistic regression coefficient interpretation>

Variable	coefficient
Age	0.0052351163
SexM	0.1778608891
ChestPainTypeATA	-0.3628929574
ChestPainTypeNAP	-0.3092947489
ChestPAinTypeTA	-0.1072256668
RestingBP	
Cholesterol	-0.0001854224
FastingBS	0.1464411883
RestingECGNormal	-0.0062603595
RestingECGST	
MaxHR	-0.0014925904
ExerciseAnginaY	0.2779756514

By using the lasso logistic regression, we can interpret our result easily. For the interpretation, the covariates CheatPainTypeATA, ChestPainTypeNAP and ExcerciseAnginaY are important in predicting heart disease. On the other hand, the covariates RestingBP and RestingECGST are not important in predicting heart disease.

Discussion

The most suitable prediction model is lasso logistic regression with the iterative regression imputation. It is not surprising that the iterative regression imputation has better accuracy. Since we have more observations, there is no information loss. The error rate is 0.166666. It means that the accuracy of prediction is 83.33334%. The most influential factor in predicting heart disease is Chest pain type. our future work would be conducting boosting and SVM models. Boosting is the updated version of the random forest. Therefore, the accuracy might be increased. According to lecture 29, SVM is the better choice when the classes are well separated (p. 11). In this case, SVM would have better predictive performance than logistic regression.