South University of Science and Technology

Models for Breast Cancer Prediction

Author: LIU, Run Qi

Course: MA403 Generalized Linear Models

Professor: CHEN, Xin

Date: May 22, 2020

Abstract

Objective To build accurate and reliable models to distinguish benign breast tumor from malignant one. **Methods** Boxplot and correlation plot are used to explore general features of data. Logistic regression, stepwise regression and random forest are used for building and selecting models. All the methods are implemented in R. **Results** The best logistic model (with least AIC) contains six features and its prediction accuracy is 97.56%. The random forest model contains nine features and its prediction accuracy is 98.05%. **Conclusions** Both logistic regression and random forest yield well-preformed model, and random forest model is with slightly stronger predictive power. Because of the existence of correlation, the random forest model is recommended as it is more robust and accurate in this case.
Keywords Breast cancer diagnosis and prognosis, Generalized linear model, Logistic Regression, Random forest

1. Introduction

Breast cancer has been identified as the most commonly occurring cancer amongst women and also a major cause of female cancer death all over the world [1]. The stage and accuracy of prognosis greatly influences the efficacy of treatment intervention, which is an important factor influencing the breast cancer mortality rate. Hence, accurate prognosis in patients with breast cancer plays a significant role to reduce mortality rate. Several factors including clump thickness, uniformity of cell size, and uniformity of cell shape can be used for diagnosing and prognosticating breast cancer. This study aims to build accurate and reliable models to distinguish benign breast tumor from malignant one. The methods used in the study include logistic regression and random forest.

2. Data

The dataset consists of 683 complete samples and each of them has 11 attributes: sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and class.

The attributes “sample code number” is the id number of observation and serves as an identifier, so it has been excluded in the analysis. The last attribute “class” is a binary variable denoting “malignant” or “benign” cases. The others variables indicate the levels of the factors related to breast cancer, and they take the integer values from one to ten. In the analysis, “class” is the response variable and the other variables except for “sample code number” are predictors. Thus, we have one nine predictors and one response. In total, there are 239 cases of malignancy, whereas benign cases are 444

During this study, I randomly split the data into two groups: training data and validating data. The training data set (n=478) consists of 70% of total samples and the validating data set (n=205) consists of 30% of total samples. There are 302 benign cases and 176 malignant cases in training set, and 142 benign cases and 63 malignant cases in validating set.

3. Exploratory data analysis

Graphical methods are used to explore the possible differences in the predictors between benign samples and malignant samples and to identify the variables that likely link to the recognition of breast cancer.

Boxplots in Figure 1 shows various attributes of benign cases and malignant cases. From Figure 1, we can see that the median value of various variables is much higher in Malignant cases. On the whole, the malignant cases have higher index of all variables than benign cases. Hence, each predictor may contribute to the recognition of malignant tumor. Furthermore, the dots in benign group in several boxplots indicate that they are outliers, exception cases. Therefore, using too few predictors in prediction would result in low accuracy.

Figure 2 illustrates the correlation between predictor variables, and the correlation coefficients between most of predictor pairs are greater than 0.5. Thus, the relationships between many predictors should not be ignored. Particularly, the correlation coefficient between uniformity of cell size and uniformity of cell shape is 0.91, indicating a strong relationship between them. In brief, the correlation matrix suggests that using all nine predictors to fit the linear model is not appropriate and it is unlikely to yield the best model.

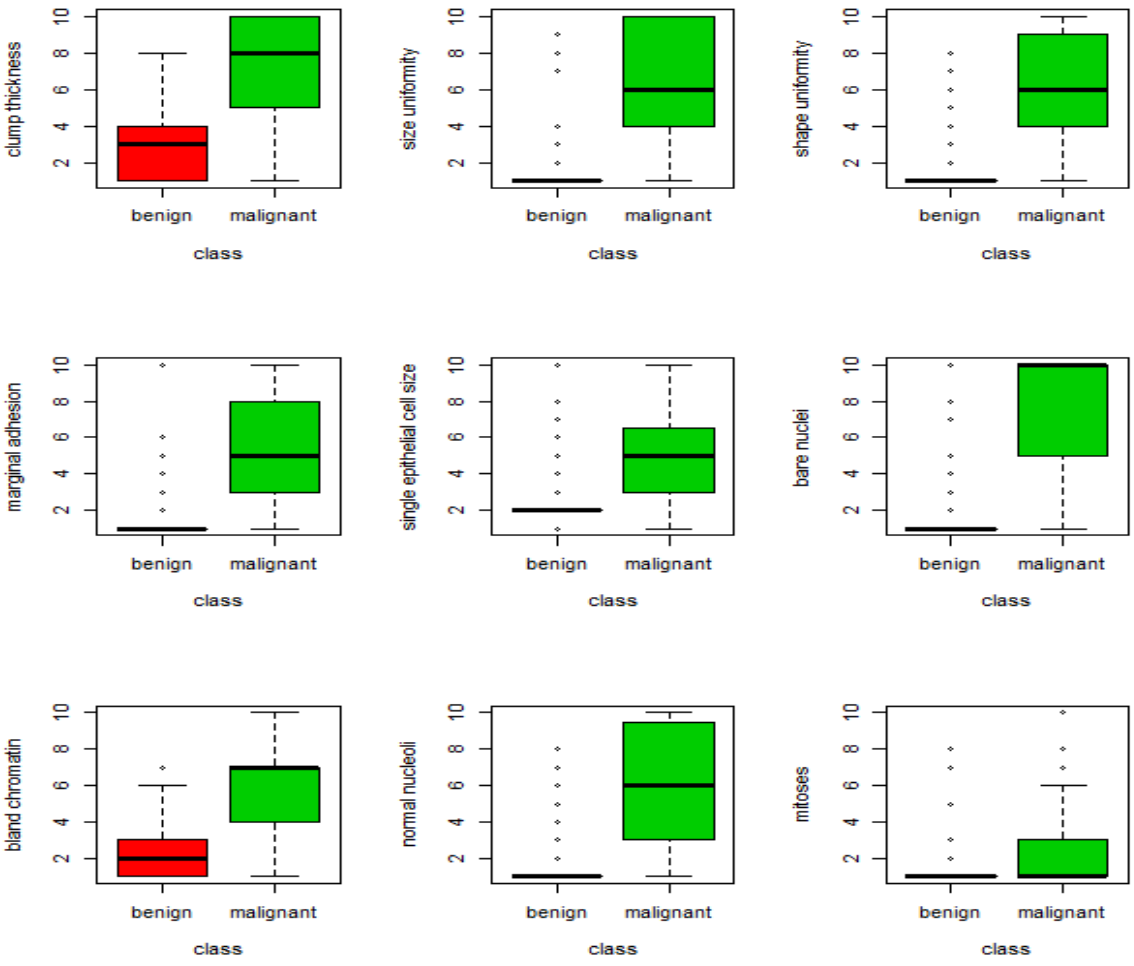

Figure 1

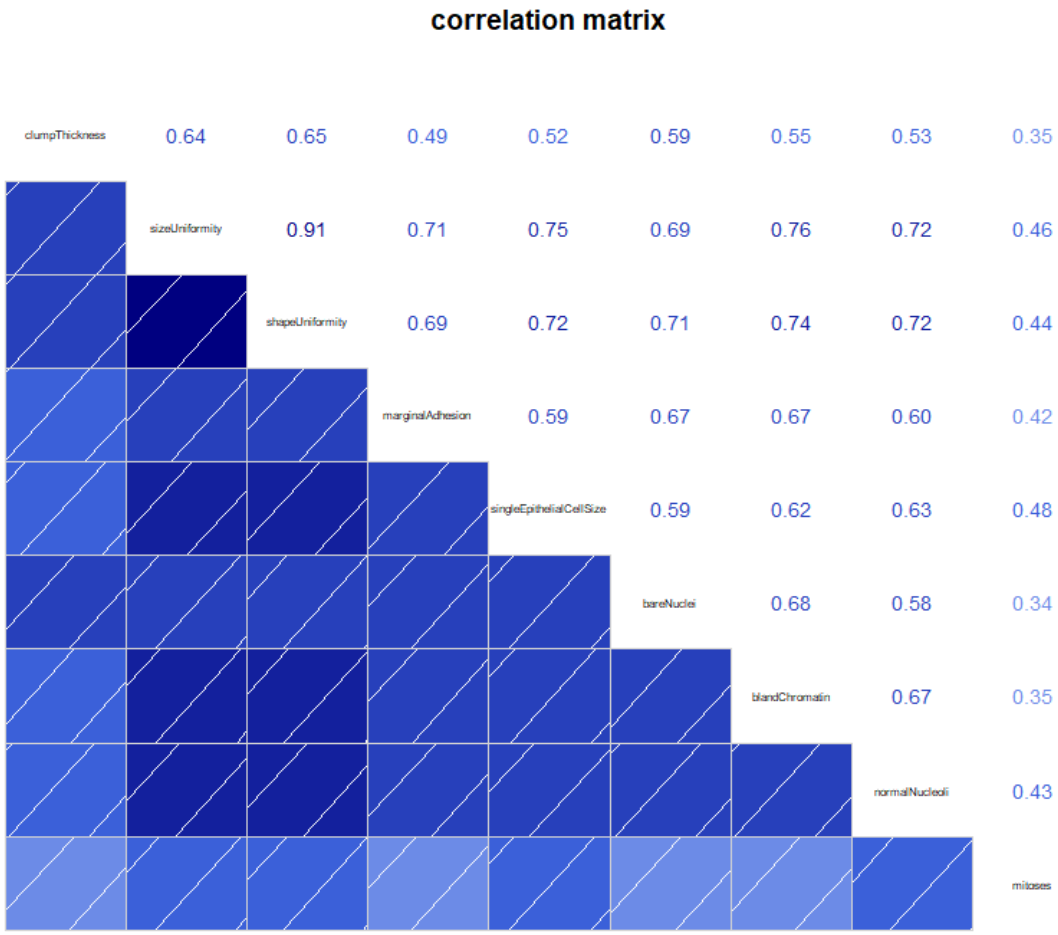

Figure 2

4. Model fitting, interpretation and evaluation

4.1 Logistic regression model

4.1.1 Model fitting and selection

Since the response variables is binary, I decide to fit a model for training data using logistic regression. Figure 3 provides the coefficients information of the full model with nine predictor variables. Under 0.1 level, only the attributes “clump thickness”, “marginal adhesion”, “bare nuclei”, “normal nucleoli”, and “mitoses” are significant. Therefore, there are some variables not important in predicting the class of tumor and the full model may not be the best model. Also, the AIC value of the full model is 100.1. Then, I fit a new model for training data with all significant variables. Figure 3 provides the coefficients information of the reduced model with six predictor variables. In this model, all the variables make a significant contribution to prediction. Also, the reduced model has a lower AIC value (95.116) than the full model, indicating that the reduced model is a better model. Chi-square test is used to compare two models, and the result (p=0.7864) suggests that the reduced model fits as well as the full model with nine

predictors. Hence, the simpler model is preferred.
Moreover, I apply the stepwise selection to find the best model with the smallest AIC. The final model of stepwise selection is the same as the reduced model, indicating that the reduced model with variables “clump thickness”, “marginal adhesion”, “bare nuclei”, “normal nucleoli”, and “mitoses” are the best logistic model for predicting the malignancy of tumor. Again, the features “uniformity of cell size”, “uniformity of cell shape” and “single epithelial cell size” are relatively not important for prediction of breast cancer. Thus, the reduced model with coefficients given in Figure 4 is adopted.

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -9.68650    1.29722  -7.467 8.20e-14 ***
clumpThickness           0.48002    0.15244   3.149  0.00164 **
sizeUniformity           0.05643    0.29272   0.193  0.84714
shapeUniformity          0.13180    0.31643   0.417  0.67703
marginalAdhesion         0.40721    0.14038   2.901  0.00372 **
singleEpithelialCellSize -0.03274    0.18095  -0.181  0.85643
bareNuclei               0.44744    0.11176   4.004 6.24e-05 ***
blandChromatin           0.48257    0.19220   2.511  0.01205 *
normalNucleoli           0.23550    0.12903   1.825  0.06798 .
mitoses                  0.66184    0.28785   2.299  0.02149 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -10.0169     1.2711  -7.881 3.25e-15 ***
clumpThickness     0.5408     0.1415   3.823 0.000132 ***
marginalAdhesion   0.4433     0.1337   3.315 0.000916 ***
bareNuclei         0.4785     0.1029   4.650 3.32e-06 ***
blandChromatin     0.5428     0.1703   3.188 0.001435 **
normalNucleoli     0.2679     0.1227   2.184 0.028955 *
mitoses            0.6949     0.2798   2.483 0.013019 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4

4.1.2 Interpretation

In the logistic regression, the response being modeled is the log(odds) that “class” = “malignant”, where odds are the ratios of the probability that class is malignant to the probability that class is benign given a set of predictors values). That is, the regression coefficients give the change in log(odds) in the response for a unit change in the predictor variable, holding all other predictor variables constant. However, log(odds) are difficult to interpret. Therefore, I exponentiate the coefficients to put the results on an odds scale.

Figure 5 gives the exponentiated coefficients of reduced model. Since all exponentiated coefficients are greater than one, the odds of malignancy increases when the index of any predictor variable increases. Specifically, the odds of a malignant case are increased by a factor of 2.003 for one-unit increase in mitoses. Also, the odds of a malignant case are increased by a factor of about 1.7 for one-unit increase in clump thickness or in bland chromatin.

```
exp(coef(fit.step))
   (Intercept)   clumpThickness marginalAdhesion       bareNuclei   blandChromatin
  4.463983e-05     1.717432e+00     1.557802e+00     1.613669e+00     1.720837e+00
normalNucleoli          mitoses
  1.307254e+00     2.003451e+00
```

Figure 5

4.1.3 Evaluation of the model performance

After adopting the reduced model, I assess its predictive power using the validating data. The confusion matrix is given in Figure 6. The prediction accuracy $\left(\left(\frac{number\ of\ true\ positive+number\ of\ true\ negative}{total\ number}\right)\times 100\%\right)$ is 97.56%. Model sensitivity $\left(\left(\frac{number\ of\ true\ positive}{number\ of\ true\ positive+number\ of\ false\ negative}\right)\times 100\%\right)$ is 95.24%. Model specificity $\left(\left(\frac{number\ of\ true\ negative}{number\ of\ true\ negative+number\ of\ false\ positive}\right)\times 100\%\right)$ is 98.59%. The precision $\left(\left(\frac{number\ of\ true\ positive}{number\ of\ true\ positive+number\ of\ false\ positive}\right)\times 100\%\right)$ is 96.77%.Since the prediction accuracy, sensitivity, specificity and precision of this model are all greater than 95%, this model is reasonable for breast cancer prediction.

```
           predict
actual      benign malignant
  benign       140         2
  malignant      3        60
```

Figure 6

4.2 Random forest model

4.2.1 Model fitting

Random forest is a flexible, easy to use machine learning algorithm that produces a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity [2]. Because of these strengths, I also apply the random forest algorithm to build a model for training data with R.

In the random forest approach, many decision trees are created and every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. Then, a new observation is fed into all the trees and taking a majority vote for each classification model.

Create random forest with R package "randomForest" with 478 training data. The number of decision trees in the forest is 500 and the number of features used as potential candidates for each split is 3. There are some cases not used while building the trees, so an OOB (out-of-bag) error estimate is made for these cases. In this model, the OOB estimate of error rate is 2.93%.

4.2.2 Interpretation

The importance of nine variables is measured based on the decrease of Gini impurity when a variable is chosen to split a node, as shown in Figure 7. The variable with a higher mean decrease of Gini is more important. Mean decrease of Gini suggests that, in this model, the uniformity of cell size is the most important feature for prediction, following by the uniformity of cell shape and bare nuclei. On the other hand, mitoses, marginal adhesion, and clump thickness are relatively not important in random forest model. The interesting thing is, two most important features, uniformity of cell size and uniformity of cell shape are not significant in logistic regression model. The correlation between variables may account for this.

```
> importance(fit.forest,type=2)
                         MeanDecreaseGini
clumpThickness                   9.656233
sizeUniformity                  60.644584
shapeUniformity                 46.361794
marginalAdhesion                 7.776020
singleEpithelialCellSize        16.070356
bareNuclei                      35.501246
blandChromatin                  27.148385
normalNucleoli                  16.283021
mitoses                          2.229869
```

Figure 7

4.2.3 Evaluation of the model performance

I assess the predictive power of random forest model using the validating data. The confusion matrix is given in Figure 8. The prediction accuracy $\left(\left(\frac{number\ of\ true\ positive+number\ of\ true\ negative}{total\ number}\right)\times 100\%\right)$ is 98.05%. Model sensitivity $\left(\left(\frac{number\ of\ true\ positive}{number\ of\ true\ positive+number\ of\ false\ negative}\right)\times 100\%\right)$ is 96.83%. Model specificity $\left(\left(\frac{number\ of\ true\ negative}{number\ of\ true\ negative+number\ of\ false\ positive}\right)\times 100\%\right)$ is 98.59%. The precision $\left(\left(\frac{number\ of\ true\ positive}{number\ of\ true\ positive+number\ of\ false\ positive}\right)\times 100\%\right)$ is 96.83%. Since the prediction accuracy, sensitivity, specificity and precision of this model are all greater than 95%, this model is also accurate for breast cancer prediction.

```
> forest.perf
           predict
actual      benign malignant
  benign       140         2
  malignant      2        61
```

Figure 8

5. Discussion and conclusions

5.1 Discussion

In this report, I have proposed two useful models for breast cancer diagnosis and prognosis, one is the logistic regression model with six predictors variables, and the other is the random forest model. The question is, which one is better in this case?
First, two models have the same specificity in predicting the validating data, 98.59%. That is, both models preform well in predicting true benign cases. Second, as for the prediction accuracy, sensitivity and precision, both models still preform well in predicting the validating data, suggesting that they both have strong predictive power. Also, the random forest model has the slightly greater values than the logistic model in the prediction accuracy, sensitivity and precision. That is, the random forest performs slightly better than the logistic model. Particularly, the random forest model performs better in predicting the malignancy cases, which is important in medical diagnosis. Since the differences in predictive power between two models are not very large, using

different validating data to assess two models can better justify the conclusion. Furthermore, because of the correlations between predictor variables, a linear model is not the best choice to fit data. On the other hand, random forest can give more accurate and robust prediction when correlated features exist. In this way, the random forest model is more practical than the logistic model in breast cancer prediction.

5.2 Conclusions

In summary, the target of this study is to build accurate and reliable models for diagnosis and prognosis breast cancer. To achieve this target, I first used graphical methods to explore the general features of data. On the whole, the malignant cases have higher index of all variables than benign cases. Also, the correlations between some features should not be ignored. Then, I split data into training data and validating data and used logistic regression and random forest to fit models. The full logistic regression model with nine predictors is not the best model and the model with smallest AIC is considered as the best logistic model. The best logistic model is with six variables: "clump thickness", "marginal adhesion", "bare nuclei", "normal nucleoli", and "mitoses". Then, I interpret the coefficients of the model and assess the predictive power of this model with validating data. The results suggest the model performs well in prediction of breast cancer. Random forest is also used to fit the model. The predictive power of random forest model is also very strong. Finally, I compared the logistic model and the random forest model. Although the performance of two models are similar in prediction of validating data, the random forest model is more favorable due to it accuracy and robustness.

6. References

[1] https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics
[2] https://builtin.com/data-science/random-forest-algorithm

7.Appendix

R codes:

```
#import data, label the class and name
the columns
library(ggplot2)
library(easyGgplot2)
cancer <- read.csv("C:/Users/ADMIN/
Desktop/breast-cancer-wisconsin.txt",
header=FALSE)
View(cancer)
cancer$V11[cancer$V11==2]=0
cancer$V11[cancer$V11==4]=1
cancer$V11=factor(cancer$V11,levels=c(0,
table(cancer$V11)

names(cancer)=c("sampleCodeNumber","clum
               "shapeUniformity","margin
               "bareNuclei","blandChroma
attach(cancer)

table(cancer$class)
##Data exploration
par(mfrow=c(3,3))
boxplot(clumpThickness~class,col=c(2,3),
thickness")
boxplot(sizeUniformity~class,col=c(2,3),
uniformity")
boxplot(shapeUniformity~class,col=c(2,3)
uniformity")
boxplot(marginalAdhesion~class,col=c(2,3
adhesion")


#Interpreting Model Parameters
coef(fit.step)
exp(coef(fit.step))

#assess the model
prob=predict(fit.step,
cancer.validate,type="response")
pred=factor(prob>0.5,levels=c(FALSE,TRUE
logit.perf=table(cancer.validate$class,p
logit.perf
```

```
###Random forest
library(randomForest)
set.seed(666)
fit.forest=randomForest(class~.,data=can
c(1)],
                        na.action =
na.roughfix,importance=TRUE)
fit.forest
importance(fit.forest,type=2)
##assess the model


forest.pred=predict(fit.forest,cancer.va
forest.perf=table(cancer.validate$class,
forest.perf
```