**Topic: Prediction of breast cancer survival**

Yuxuan Wang, yw3601

**I. Introduction**

According to the Centers for Disease Control and Prevention, breast cancer is the most common cancer among women, affecting 2.1 million women each year. And it causes the most cancer-related deaths among women. A report showed that there were 627,000 women died of breast cancer in 2018. That’s why we want to study this topic. If we could find a proper model to predict the survival rate of breast cancer patients, it could help doctors to make decisions and may save patients’ lives.

Through literature review, we know that the most important part of clinical decision-making for cancer patients is to accurately estimate prognosis and survival. For breast cancer patients with the same disease stage and clinical characteristics, treatment response and overall survival may be different. The data set we use is from the International Society for molecular Classification of Breast Cancer (METABRIC) database, including 31 clinical attributes, m-RNA levels z-score for 331 genes, and mutation in 175 genes for 1904 breast cancer patients. The use of machine learning techniques on the data allows estimates of patient survival rate, a better understanding of cancer prognosis and outcomes, and can prevent unnecessary surgeries and treatment procedures. For this aim, after comparing several kinds of models, we finally choose random forest to train our model.

**II. Data processing**

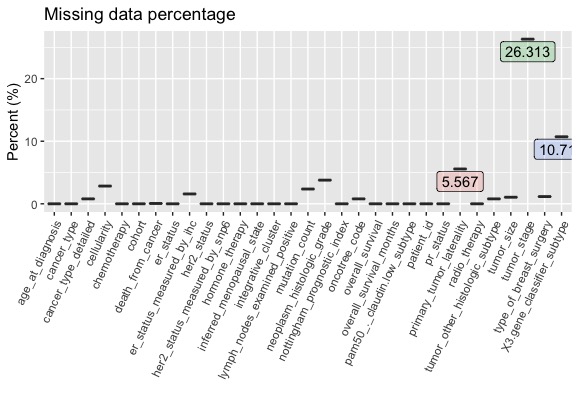
In our study, we use the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) data set. The data set contains 31 clinical attributes, m-RNA levels z-score for 331 genes, and mutation in 175 genes for 1904 breast cancer patients. Firstly, we just use the 31 clinical variables in the data set, which could be categorized into 5 groups by definition as listed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tumor info.** | **Cancer type** | **General info.** | **Treatment** | **Patient status** |
| cellularity | cancer\_type | patient\_id | typeofbreast\_surgery | overallsurvivalmonths |
| pam50+claudin-low\_subtype | cancertypedetailed | ageatdiagnosis | chemotherapy | overall\_survival |
| erstatusmeasuredbyihc | tumorotherhistologic\_subtype | inferredmenopausalstate | hormone\_therapy | deathfromcancer |
| er\_status | integrative\_cluster | primarytumorlaterality | radio\_therapy |  |
| neoplasmhistologicgrade | lymphnodesexamined\_positive |  |  |  |
| her2statusmeasuredbysnp6 | nottinghamprognosticindex |  |  |  |
| her2\_status | oncotree\_code |  |  |  |
| mutation\_count | tumor\_stage |  |  |  |
| pr\_status |  |  |  |  |
| 3-geneclassifiersubtype |  |  |  |  |
| tumor\_size |  |  |  |  |

Table1. Clinical variables category

Variables in the first column are information of tumor cells, for example, tumor size measured by imaging techniques. The second column includes variables about cancer type. The third column includes variables on general information of patients. The fourth column includes variables about types of treatment and the fifth column includes variables about survival situation of patients.

As we all know, the presence of too much missing value in the data set will have negative effect on precision and accuracy of the model. Generally, less than 5% missing data rate is an acceptable threshold. We calculate percentage of missing data for each variable and remove variables with missing data percentage greater than 5%. By Plot1, we decide to remove three redundant variables: *tumor\_stage, X3-geneclassifiersubtype, primary\_tumor\_laterality*.



Plot1. Missing data percentage

By analyzing definition of the variables:

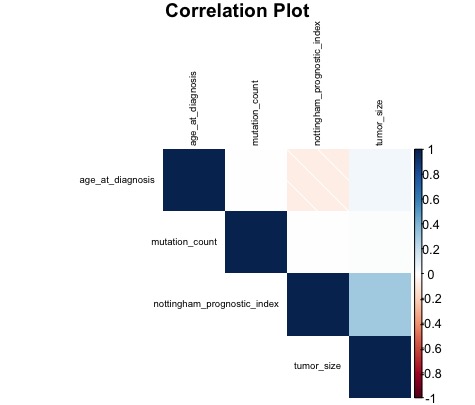
(1)In the cancer type group, variables *cancer\_type*, *cancer\_type\_detailed*, *tumor\_other\_histologic\_subtype*, *oncotree\_code* present redundant information. So we only keep one of them in our variable set.

(2)We consider *er\_status* and *er\_status\_measured\_by\_ihc* gives same information, where *her2\_status* and *her2\_status\_measured\_by\_snp6* gives same information.

(3)We consider *overall\_survival\_months* and *death\_from\_cancer* present the same information as *overall\_survival*.

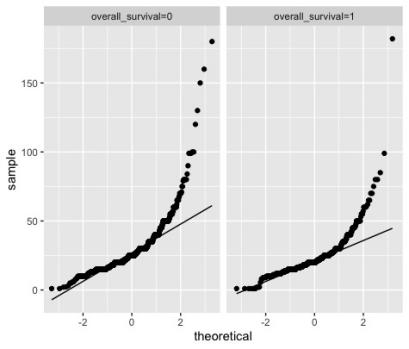
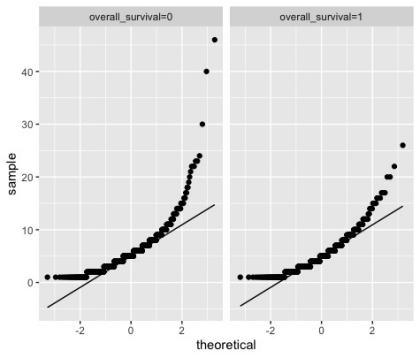
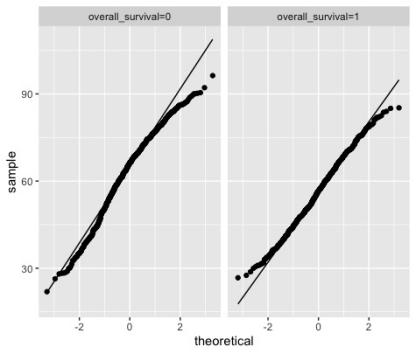
(4)*Patient\_id* and *cohort* do not need to be concluded in variable set.

After removing redundant variables, there are 19 variables kept in our model. For categorical variables *tumor\_other\_histologic\_subtype* and *pam50\_.\_claudin.low\_subtype*, we remove data that *tumor\_other\_histologic\_subtype* is "Metaplastic" and *pam50\_.\_claudin.low\_subtype* is "NC" because of the too low sample size of these two categories. Finally, we drop missing values and get the pre-training data set.



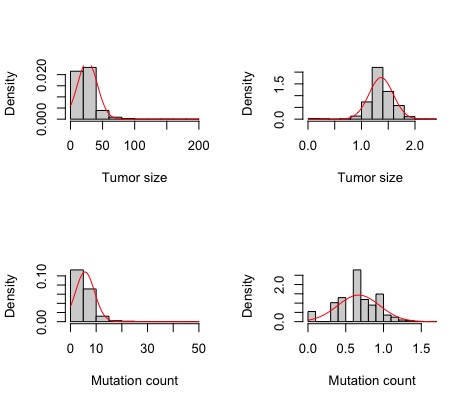
Plot2. Correlation plot

Plot2 shows all pairwise correlation between all numerical variables in our data set, which indicates that *nottingham\_prognostic\_index* and *tumor\_size* are positively related, while *age\_at\_diagnosis* and *nottingham\_prognostic\_index* are negatively related.



Plot3. Q-Q plots of age\_at\_diagnosis, mutation\_count, tumor\_size

By Q-Q plot of the three continuous numerical variables, we decide to do log transformation to *mutation\_count* and *tumor\_size*.



Plot4. Log transformation

We can see that after log transformation, the distribution of *mutation\_count* and *tumor\_size* become normal. So the transformation makes sense.

**III. Methods**

We randomly choose 90% of the total sample size as the training set, and the other data as test set. Then we try several linear models and nonlinear models. Among them, linear models include logistic, lasso, ridge and elastic net, while nonlinear models include decision tree, random forest, xgboost. And we analyze the interpretability of each model.

For linear models: (1) As a kind of regression model, logistic model is easy to understand and interpret. But at the same time, it is more sensitive to multicollinearity and always have lower precision because it’s too simple. (2) Lasso is a good method to do variable selection because it can compress coefficients of variables, while ridge cannot compress the coefficients to zero. And elastic net is the combination of lasso and ridge.

For nonlinear models: (1) There are two kinds of decision trees that use different criteria of split - ID3 using information entropy, and CART using gini index. (2) The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. As a result, it is more robust. (3) XGboost is a scalable, distributed gradient-boosted decision tree machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

**IV. Result**

First of all, we analyze the importance of variables by using different models to have a preliminary understanding.

1. For logistic model, we can get a table that describes the importance of variables. From the table, we conclude that variables *age\_at\_diagnosis*, *hormone\_therapy1*, *inferred\_menopausal\_statePre*, *nottingham\_prognostic\_index*, *tumor\_size\_T* are relatively of great importance to the survival rate.
2. As we all know, lasso and ridge are two special cases of elastic net. Compared to ridge, lasso and elastic net are relatively better methods to do variable selection because they can compress coefficients of unimportant variables to zero. By lasso and elastic net, variables *age\_at\_diagnosis*, *hormone\_therapy1*, *inferred\_menopausal\_statePre*, *nottingham\_prognostic\_index*, *tumor\_size\_T* are of great importance to the survival rate. But by ridge, we do not know which variable is more important.
3. For decision tree models, we can conclude that *age\_at\_diagnosis*, *nottingham\_prognostic\_index*, *pam50\_.\_claudin.low\_subtype*, *integrative\_cluster*, *inferred\_menopausal\_state*, *her2\_status* are the most important variables that listed in decreasing importance. What is more, *age\_at\_diagnosis*, *nottingham\_prognostic\_index*, *integrative\_cluster*, *inferred\_menopausal\_state* are in the top 5 important variables both in the two models.
4. By random forest method, we can conclude that *age\_at\_diagnosis*, *nottingham\_prognostic\_index*, *integrative\_cluster*, *tumor\_size\_T*, *mutation\_count\_T* are the most important variables for survival rate*.*
5. By XGboost, *age\_at\_diagnosis*, *nottingham\_prognostic\_index*, *mutation\_count\_T, tumor\_size\_T*, *pam50\_.\_claudin.low\_subtypeLumA* are the most important variables*.*

In conclusion, for all the methods, we can primarily conclude that *age\_at\_diagnosis, nottingham\_prognostic\_index, tumor\_size\_T* are the most important variables.

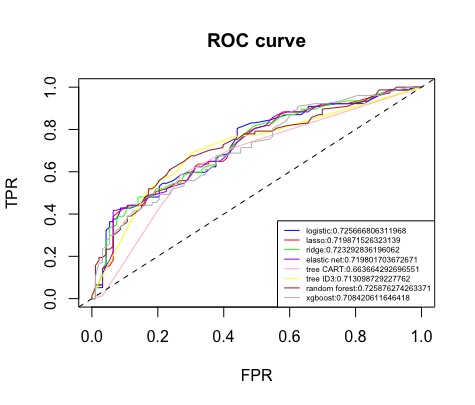
Then we will observe test error for different models by using accuracy, precision, recall and F1 score, the results are as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | accuracy score | precision score | recall score | F1-score |
| logistic\_model | 0.6705882 | 0.6728972 | 0.7741935 | 0.72 |
| lasso model | 0.6705882 | 0.6637168 | 0.8064516 | 0.7281553 |
| Ridge model | 0.6588235 | 0.6666667 | 0.7526882 | 0.7070707 |
| elastic net model | 0.6705882 | 0.6637168 | 0.8064516 | 0.7281553 |
| decision tree CART | 0.641176 | 0.6454545 | 0.7634409 | 0.6995074 |
| decision tree ID3 | 0.6882353 | 0.6886792 | 0.7849462 | 0.7336683 |
| random forest | 0.6823529 | 0.6756757 | 0.8064516 | 0.7352941 |
| xgboost | 0.6470588 | 0.6774194 | 0.6774194 | 0.6774194 |

Table2. Scores for models

From the table, we can primarily conclude that decision tree ID3 and random forest are the most robust models for they have higher accuracy score. The precision score (TP/(TP+FP)) represents the percentage of actual positive samples in total positive results. It should be considered as an important criterion since the cost of predicting negative samples into positive samples (FP) is so large, which means that the patient is estimated to death but actually he could survive. What is more, the recall score (TP/(TP+FN)) represents the percentage of actual positive samples in total positive samples and it should also be considered as an important criterion when the cost of predicting false negative is large, which means that the patient is estimated to survive but die in reality. It’s generally admitted that if a patient is predicted to die, medical institutions should pay more attention to them and attempt to make strategy to increase survival rate. And if a patient is predicted to survive but not surviving from the cancer, medical institutions may pay less attention to them and miss the best time for treatment. As a result, recall score should be an important criterion to consider.

By table2, the recall score and F1-score of random forest are 0.8064516 and 0.7352941 respectively, which is higher than that in decision tree ID3. So, we consider random forest as the best model to predict survival rate for breast cancer patients.



Plot5. ROC curve for models

ROC curve is always used to judge the accuracy of model predictions, in which the x-axis is specificity and the y-axis is sensitivity. The closer the curve locates to the upper left corner, the larger the area below the curve, which means a higher accuracy score. From plot4, we can see that the curve of random forest is the closest to the upper left corner, which is consistent with our previous conclusion.

After choosing random forest as our model, we could then analyze the importance of variables by using MeanDecreaseAccuracy and MeanDecreaseGini as follows:

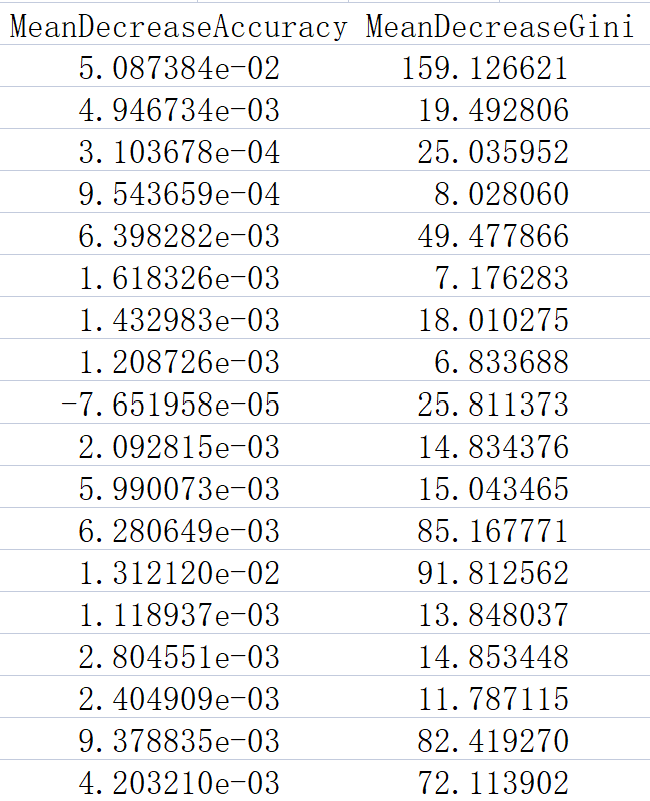
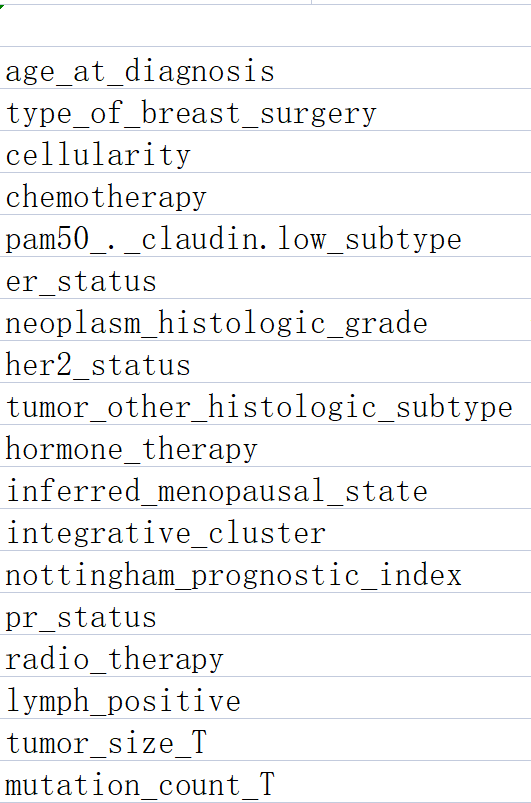


Table3. importance of variables in random forest

As we all know, variables with higher index means higher importance in the model. From table3, we can conclude that *age\_at\_diagnosis* is the most important variable in our model with Gini index 159.126. So we can conclude that a*ge\_at\_diagnosis* has significant impact on survival and should be paid much attention to. What is more, the top5 important variables for survival are: *age\_at\_diagnosis, nottingham\_prognostic\_index*, *integrative\_cluster*, *tumor\_size\_T*, *mutation\_count\_T.*

**V. Conclusion and improvement**

In conclusion, the best-performed model we choose to predict survival rate is random forest, with accuracy score 0.6823529. And the top5 important variables are things that should be mostly concerned about in real-life decision-making by doctors: A*ge\_at\_diagnosis* represents age of *the* patient at diagnosis time, which is consistent to the common sense that older patients may get more basic illnesses and have lower possibility to survive. *Nottingham\_prognostic\_index* is calculated using three pathological criteria - the size of the tumor; the number of involved lymph nodes; and the grade of the tumor, which represents information of tumor cells. *Integrative\_cluster* represents molecular subtype of the cancer based on some gene expression. *Tumor\_size\_T* represents tumor size measured by imaging techniques. *Mutation\_count\_T* represents number of gene that has relevant mutations. These four variables are truly need attention since they include fundamental message of tumor.

The accuracy score of random forest still could be improved. Except for clinical attributes, the data set also provide genetic attributes, which contains m-RNA levels z-score for 331 genes, and mutation for 175 genes. To further improve our model, we could use PCA to do variable selection for the genes, add more variables to train model and do the prediction.