Credit Card Customer Default Risk Forecast Report

Mia Lai

I. Experimental purpose

1. The data set of credit card customers in Taiwan from April 2005 to September 2005 on UCI website was used to preprocess the data, including data cleaning and data conversion, etc., and master the relevant principles and operations;
2. The Apriori algorithm is used to conduct correlation analysis on the processed data mart to find the variables strongly related to customer default risk, to judge which types of customers are more likely to default;
3. Random forest algorithm is used to model the data set to predict the default customers in the next month.
4. The confusion matrix and ROC curve were used to evaluate the prediction effect.

II. Experimental principle

1. Data preprocessing

In most cases, the original data should not be directly used for modeling, and it is necessary to preprocess the data in various aspects before entering the modeling process. In general, data preprocessing includes:

(1) Data cleansing refers to the last step of finding and correcting identifiable errors in data files, including checking data consistency, dealing with invalid and missing values, etc. Unlike questionnaire review, data cleaning after entry is generally done by computer rather than human.

(2) Data conversion, the original data is not suitable for direct modeling, and some transformations are needed so that the model can be better used directly.

2. Association rule algorithm

The association algorithm is an important class of algorithms in data mining. Its core is recursion based on the idea of two-stage frequent sets. It first finds out all the frequency sets whose item sets appear at least as frequently as the predefined minimum support. Then strong association rules are generated from the frequency set, which must satisfy minimum support and minimum confidence. Finally, the frequency set found in Step 1 is used to produce the desired rules, producing all rules that contain only the terms of the set, where each rule has only one term on the right.

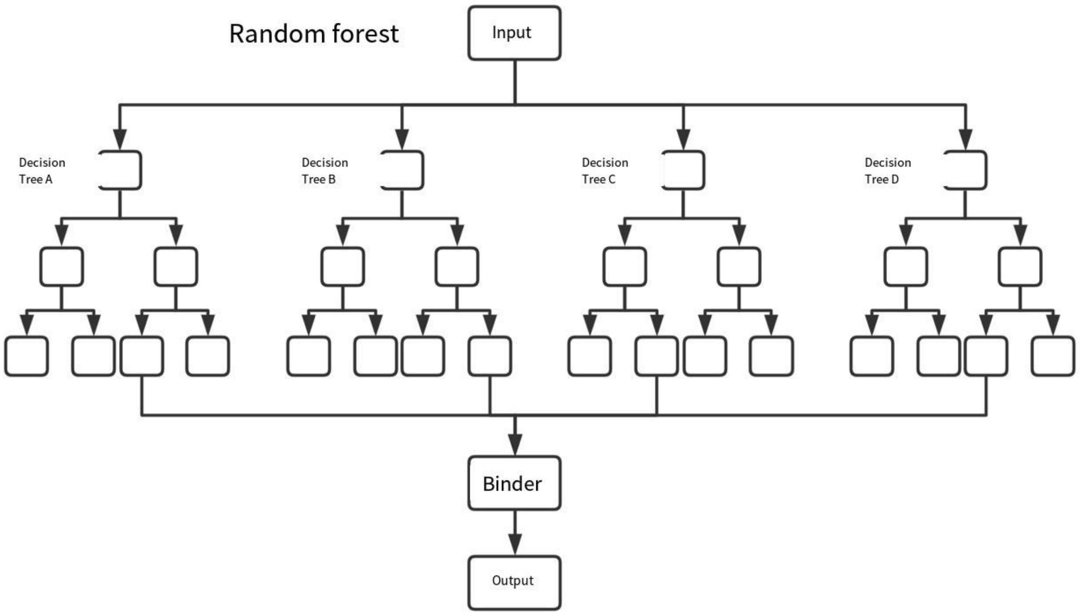
Apriori algorithm is the most used algorithm for association rules. It divides the process of discovering association rules into two steps. The first step is to iteratively retrieve all frequent item sets in transaction database 1, that is, item sets whose support degree is not lower than the threshold set by the user; In the second step, frequent item sets are used to construct rules that meet the minimum trust of users. Among them, mining or identifying all frequent item sets is the core of the algorithm, accounting for most of the calculation.

3. Random Forest algorithm

Random forest refers to a classifier that uses multiple trees to train and predict samples. It belongs to the Bagging type. By combining multiple weak classifiers, the result can be voted or averaged, so that the results of the overall model have high accuracy and generalization performance. It can achieve good results mainly due to "random" and "forest", one of which makes it resistant to overfitting and the other makes it more accurate.

Random Forest uses decision trees as weak learners; Second, based on the use of decision trees, Random Forest improves the establishment of decision trees.

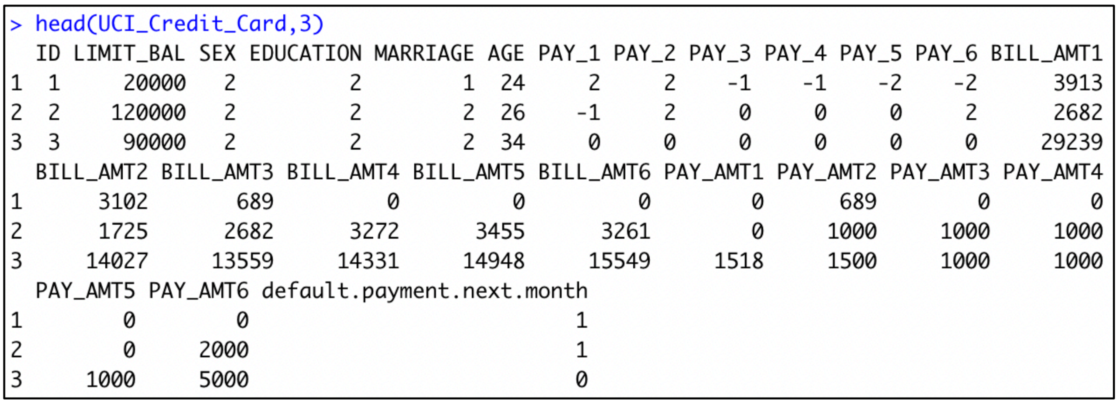
For ordinary decision trees, we will choose an optimal feature among all n sample features on the nodes to divide the left and right subtrees of the decision trees. However, random forest randomly selects some sample features on the nodes, and the number is less than n. Assume that it is $nsub$, and then select an optimal feature among the $nsub$ sample features randomly to partition the left and right subtrees of the decision tree. This further enhances the generalization ability of the model.



III. Experimental process and analysis

1. Data preprocessing

(1) Load the package and import the data set to view the data. Key variables include things like gender, age, marital status, and education;



```
> head(UCI_Credit_Card,3)
  ID LIMIT_BAL SEX EDUCATION MARRIAGE AGE PAY_1 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6 BILL_AMT1
1  1     20000   2         2        1  24     2     2    -1    -1    -2    -2      3913
2  2    120000   2         2        2  26    -1     2     0     0     0     2      2682
3  3     90000   2         2        2  34     0     0     0     0     0     0     29239
  BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4
1      3102       689         0         0         0        0      689        0        0
2      1725      2682      3272      3455      3261        0     1000     1000     1000
3     14027     13559     14331     14948     15549     1518     1500     1000     1000
  PAY_AMT5 PAY_AMT6 default.payment.next.month
1        0        0                          1
2        0     2000                          1
3     1000     5000                          0
```

The explanation of some variables is shown in the table below:

Column names	Description of Meaning
ID	Customer ID
LIMIT_BAL	A line of credit that a bank gives to a customer, including a personal line of credit and a customer's family line of credit.
SEX	Gender of the client. Note the male as 1 and the female as 2.
EDUCATION	The client's level of education. Graduate and above is listed as 1, college is listed as 2, high school is listed as 3, other is listed as 4.
MARRIAGE	The client's marital status. Married as 1, unmarried as 2, other as 3.
AGE	The age of the client.
PAY_1 to PAY_6	The six variables are monthly payments from April to September 2005
BILL_AMT1 ~ BILL_AMT6	The six variables are the monthly billing records, that is, the monthly credit card purchases, from April to September 2005.
PAY_AMT1 ~ PAY_AMT6	The six variables are monthly payment records from April to September 2005, including the amount of bills paid back and the amount deposited to the credit card.

default.payment.next.month	Default or not in the next month, default is recorded as 1, no default is recorded as 0

Additional notes:

PAY_1 ~ PAY_6: Repayment from September 2005 to April 2005

PAY_1 indicates the repayment status from September 2005; PAY_2 is the repayment in August 2005; ... ; PAY_6 is repayment as of April 2005. The same is true for the meaning of the numeric identifiers in BILL_AMT1 to BILL_AMT6 and PAY_AMT1 to PAY_AMT6.

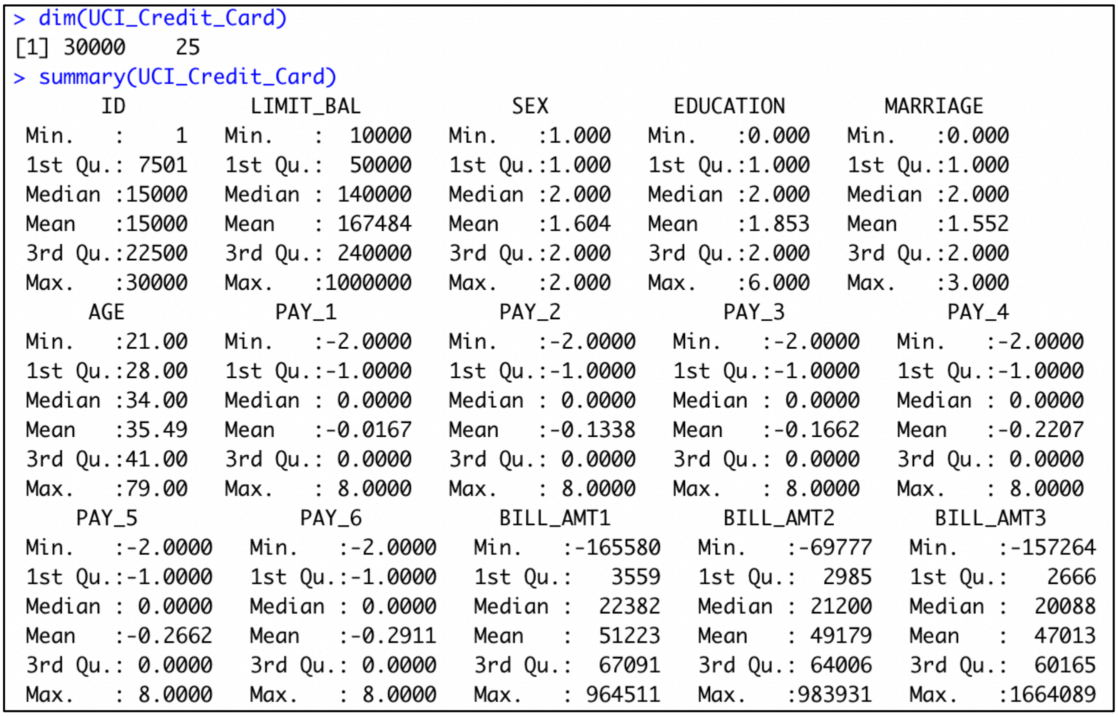
The value meanings of PAY_1 to PAY_6 are as follows:

-2 = two months in advance; -1 = one month in advance; 0 = timely return; 1 = one month late; 2 = two months late; 3 = three months late; ... ; 9 = repayment delays of nine months or more.

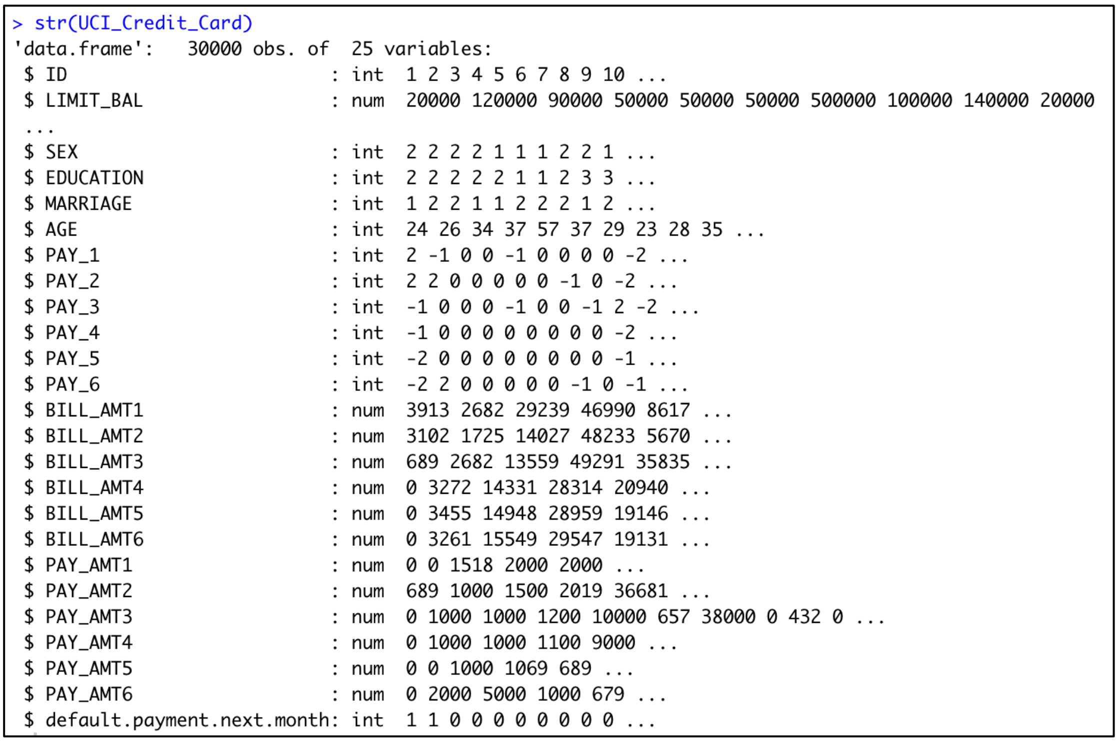
The monthly payment amount PAY_AMT cannot fall below the bank's minimum payment for the month, otherwise it is a default. If the payment amount PAY_AMT is greater than the bill amount of last month BILL_AMT, it will be regarded as timely repayment, and the remaining amount will be deposited in the credit card for next consumption; If the payment amount is less than the previous month's bill amount but higher than the minimum payment amount, it is considered as late payment.

(2) Perform descriptive line statistics on the raw data and look at each variable type;

The original data is a 30000*25 data structure, with a total of 30,000 customers and 25 related variables;

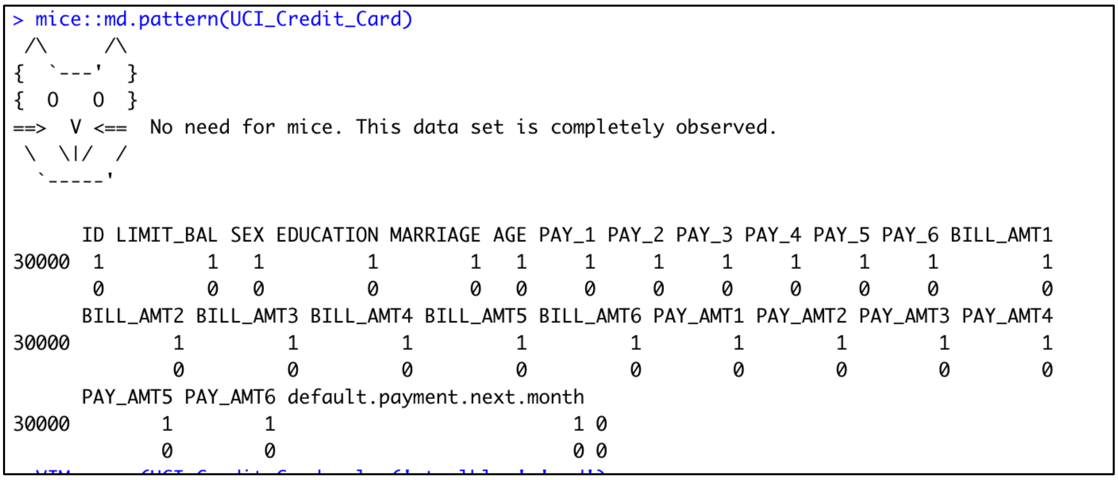


```
> dim(UCI_Credit_Card)
[1] 30000    25
> summary(UCI_Credit_Card)
       ID          LIMIT_BAL            SEX          EDUCATION        MARRIAGE    
 Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000   Min.   :0.000  
 1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :15000   Median : 140000   Median :2.000   Median :2.000   Median :2.000  
 Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853   Mean   :1.552  
 3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000   Max.   :3.000  
      AGE            PAY_1             PAY_2             PAY_3             PAY_4        
 Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000  
 1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000  
 Median :34.00   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000  
 Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338   Mean   :-0.1662   Mean   :-0.2207  
 3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000  
 Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000  
     PAY_5             PAY_6           BILL_AMT1         BILL_AMT2        BILL_AMT3      
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-165580   Min.   :-69777   Min.   :-157264  
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666  
 Median : 0.0000   Median : 0.0000   Median :  22382   Median : 21200   Median :  20088  
 Mean   :-0.2662   Mean   :-0.2911   Mean   :  51223   Mean   : 49179   Mean   :  47013  
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165  
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 964511   Max.   :983931   Max.   :1664089  
```

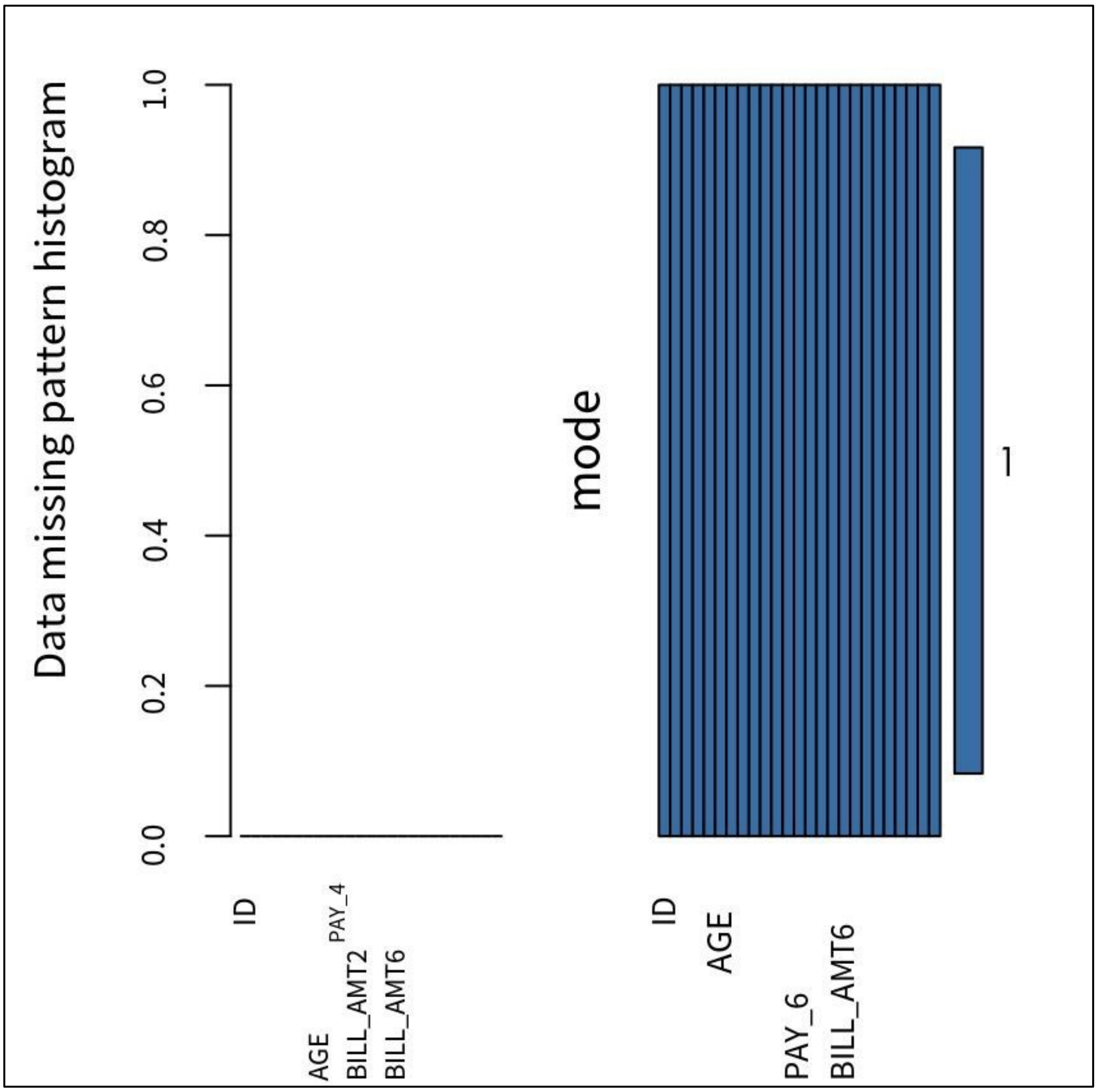


```
> str(UCI_Credit_Card)
'data.frame':	30000 obs. of  25 variables:
 $ ID                        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL                 : num  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000
 ...
 $ SEX                       : int  2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : int  2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : int  1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : int  24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_1                     : int  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2                     : int  2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3                     : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4                     : int  -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5                     : int  -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6                     : int  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1                 : num  3913 2682 29239 46990 8617 ...
 $ BILL_AMT2                 : num  3102 1725 14027 48233 5670 ...
 $ BILL_AMT3                 : num  689 2682 13559 49291 35835 ...
 $ BILL_AMT4                 : num  0 3272 14331 28314 20940 ...
 $ BILL_AMT5                 : num  0 3455 14948 28959 19146 ...
 $ BILL_AMT6                 : num  0 3261 15549 29547 19131 ...
 $ PAY_AMT1                  : num  0 0 1518 2000 2000 ...
 $ PAY_AMT2                  : num  689 1000 1500 2019 36681 ...
 $ PAY_AMT3                  : num  0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4                  : num  0 1000 1000 1100 9000 ...
 $ PAY_AMT5                  : num  0 0 1000 1069 689 ...
 $ PAY_AMT6                  : num  0 2000 5000 1000 679 ...
 $ default.payment.next.month: int  1 1 0 0 0 0 0 0 0 0 ...
```

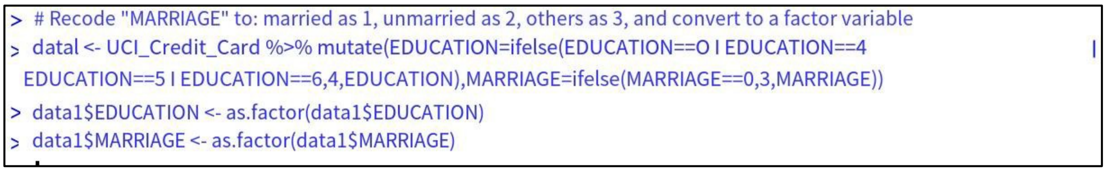
(3) Check whether there is any missing data by using the data exact pattern histogram, and there is no missing data.



```
> mice::md.pattern(UCI_Credit_Card)
 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'

      ID LIMIT_BAL SEX EDUCATION MARRIAGE AGE PAY_1 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6 BILL_AMT1
30000  1         1   1         1        1   1     1     1     1     1     1     1         1
       0         0   0         0        0   0     0     0     0     0     0     0         0
      BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4
30000         1         1         1         1         1        1        1        1        1
              0         0         0         0         0        0        0        0        0
      PAY_AMT5 PAY_AMT6 default.payment.next.month
30000        1        1                          1 0
             0        0                          0 0
```

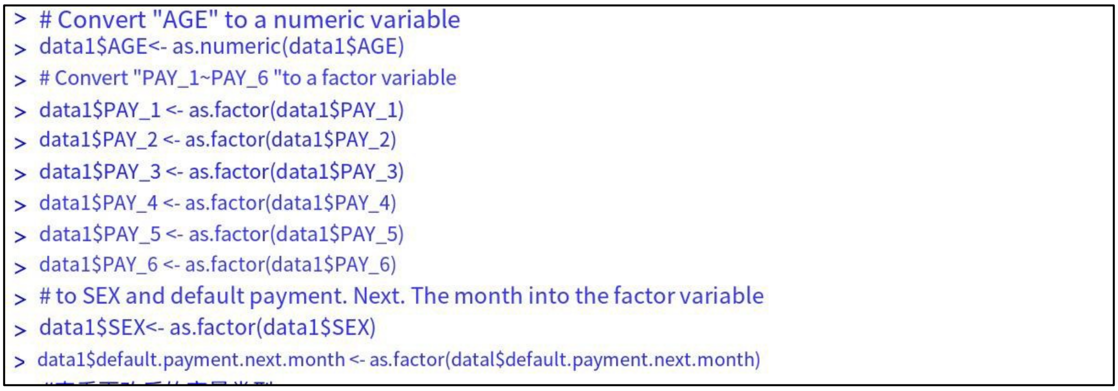

Data missing pattern histogram
1.0
0.8
0.6
0.4
0.2
0.0
ID
AGE
PAY_4
BILL_AMT2
BILL_AMT6
mode
ID
AGE
PAY_6
BILL_AMT6
1

(4) The 0, 4, 5, and 6 in the "EDUCATION" variable of the original data all represent other cases. For the convenience of grouping these codes into one class, they are recorded as follows: Graduate students and above are recorded as 1, the university is recorded as 2, high school is recorded as 3, other is recorded as 4, and converted into factor variables, because "MARRIAGE" variable 0 code corresponding data can rarely be classified into other classes, recoded as married as 1, unmarried as 2, other as 3, and converted into factor variables;

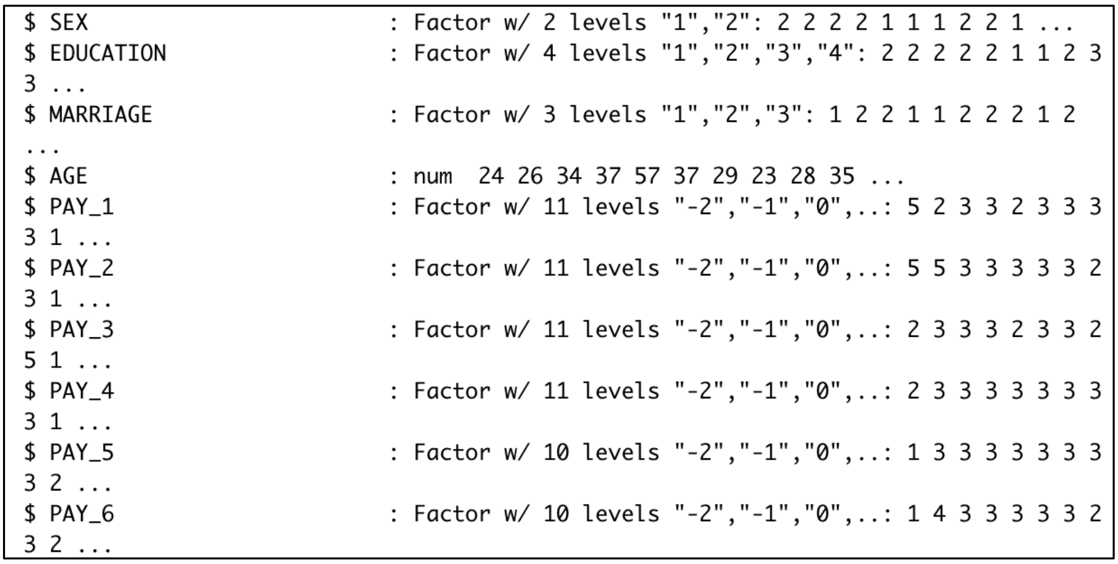


```
> # Recode "MARRIAGE" to: married as 1, unmarried as 2, others as 3, and convert to a factor variable
> datal <- UCI_Credit_Card %>% mutate(EDUCATION=ifelse(EDUCATION==O I EDUCATION==4                    I
 EDUCATION==5 I EDUCATION==6,4,EDUCATION),MARRIAGE=ifelse(MARRIAGE==0,3,MARRIAGE))
> data1$EDUCATION <- as.factor(data1$EDUCATION)
> data1$MARRIAGE <- as.factor(data1$MARRIAGE)
```

(5) The "AGE" into a numerical variable, PAY_1~PAY_6, SEX and ‘default.payment.next.month’ is converted to a factor variable, and check whether the conversion is successful;

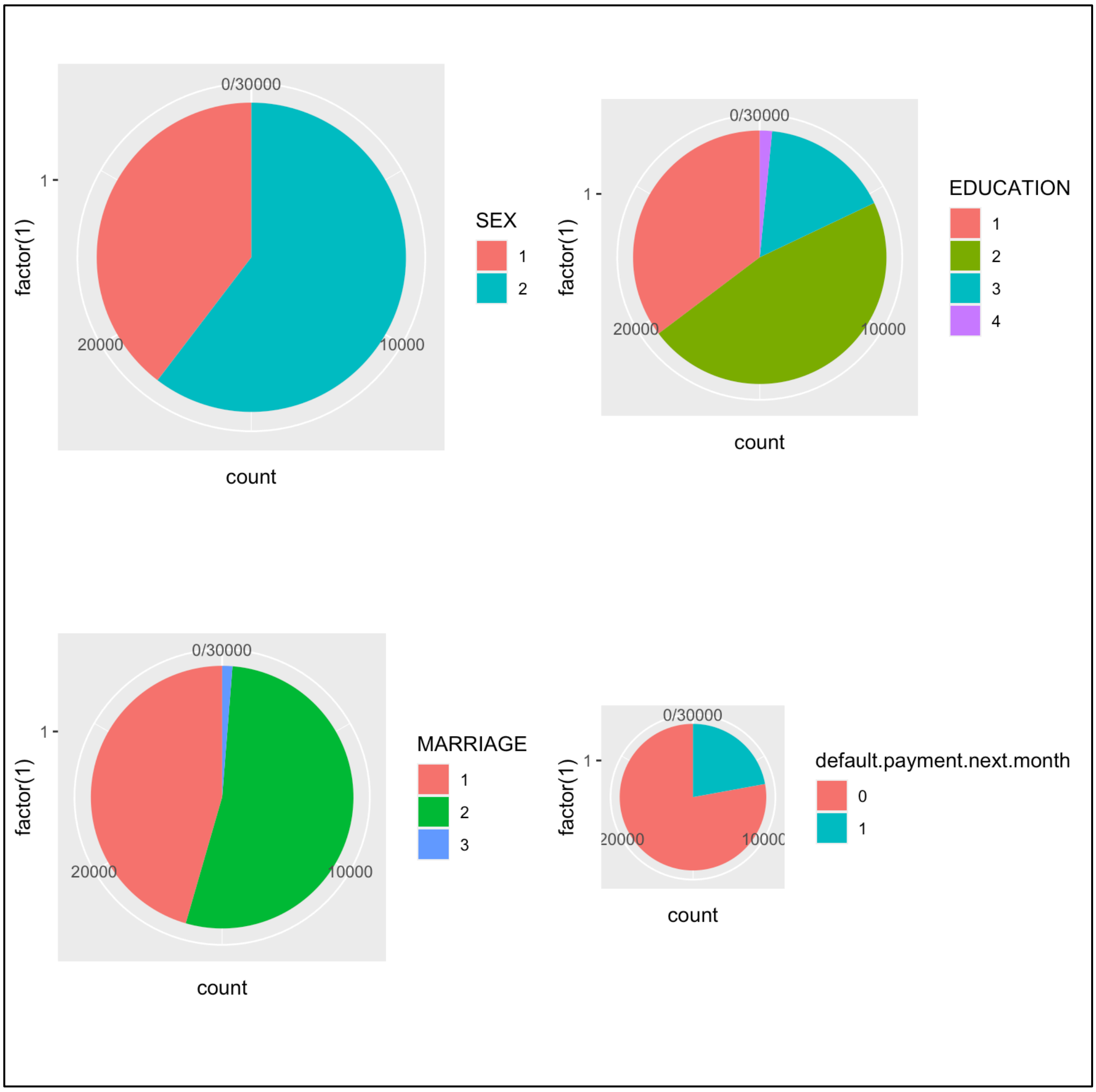


```
> # Convert "AGE" to a numeric variable
> data1$AGE<- as.numeric(data1$AGE)
> # Convert "PAY_1~PAY_6 "to a factor variable
> data1$PAY_1 <- as.factor(data1$PAY_1)
> data1$PAY_2 <- as.factor(data1$PAY_2)
> data1$PAY_3 <- as.factor(data1$PAY_3)
> data1$PAY_4 <- as.factor(data1$PAY_4)
> data1$PAY_5 <- as.factor(data1$PAY_5)
> data1$PAY_6 <- as.factor(data1$PAY_6)
> # to SEX and default payment. Next. The month into the factor variable
> data1$SEX<- as.factor(data1$SEX)
> data1$default.payment.next.month <- as.factor(datal$default.payment.next.month)
```



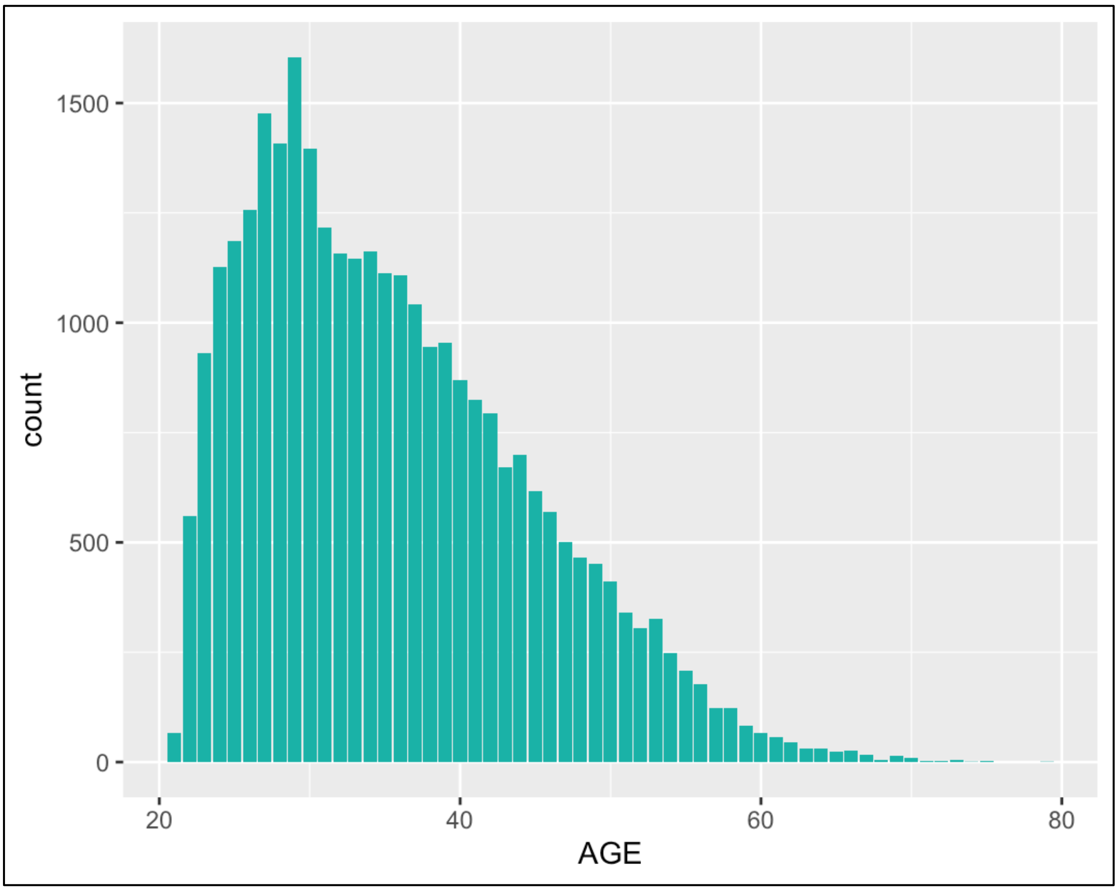
```
$ SEX                        : Factor w/ 2 levels "1","2": 2 2 2 2 1 1 1 2 2 1 ...
$ EDUCATION                  : Factor w/ 4 levels "1","2","3","4": 2 2 2 2 2 1 1 2 3
3 ...
$ MARRIAGE                   : Factor w/ 3 levels "1","2","3": 1 2 2 1 1 2 2 2 1 2
...
$ AGE                        : num  24 26 34 37 57 37 29 23 28 35 ...
$ PAY_1                      : Factor w/ 11 levels "-2","-1","0",..: 5 2 3 3 2 3 3 3
3 1 ...
$ PAY_2                      : Factor w/ 11 levels "-2","-1","0",..: 5 5 3 3 3 3 3 2
3 1 ...
$ PAY_3                      : Factor w/ 11 levels "-2","-1","0",..: 2 3 3 3 2 3 3 2
5 1 ...
$ PAY_4                      : Factor w/ 11 levels "-2","-1","0",..: 2 3 3 3 3 3 3 3
3 1 ...
$ PAY_5                      : Factor w/ 10 levels "-2","-1","0",..: 1 3 3 3 3 3 3 3
3 2 ...
$ PAY_6                      : Factor w/ 10 levels "-2","-1","0",..: 1 4 3 3 3 3 3 2
3 2 ...
```

(6) Visualize the converted data. Look at the distribution of factor gender, marital status, education level, and default situation, numerical age distribution, gender, marital status, and education level distribution under different default situations, age, and credit limit distribution, and give a brief description;



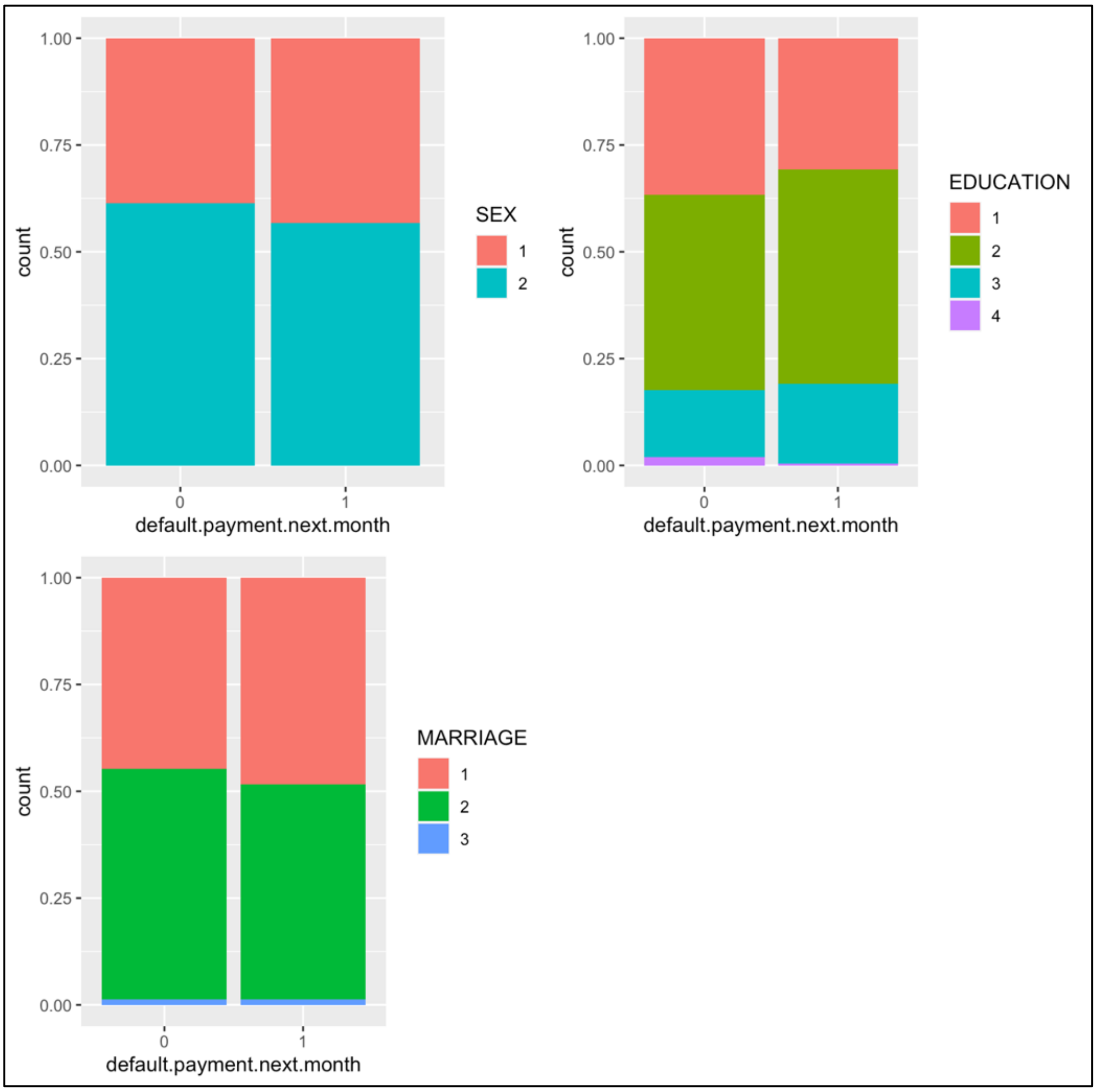
Customer gender, education, marital status, and default distribution map

As can be seen from the graph above, more single people use credit cards than married ones; The number of bachelor's degree holders using credit cards is the highest; More women than men use them; From the distribution of dependent variables, it can be seen that there are far more performing customers than defaulting customer



Age distribution of customers

As can be seen from the age, the majority of customers using credit cards are young people, with an intensive age distribution between 20 and 40 years old, and the peak age is around 30 years old. With the increase of age, the number of customers using credit cards decreases sharply, showing a right-sided distribution.



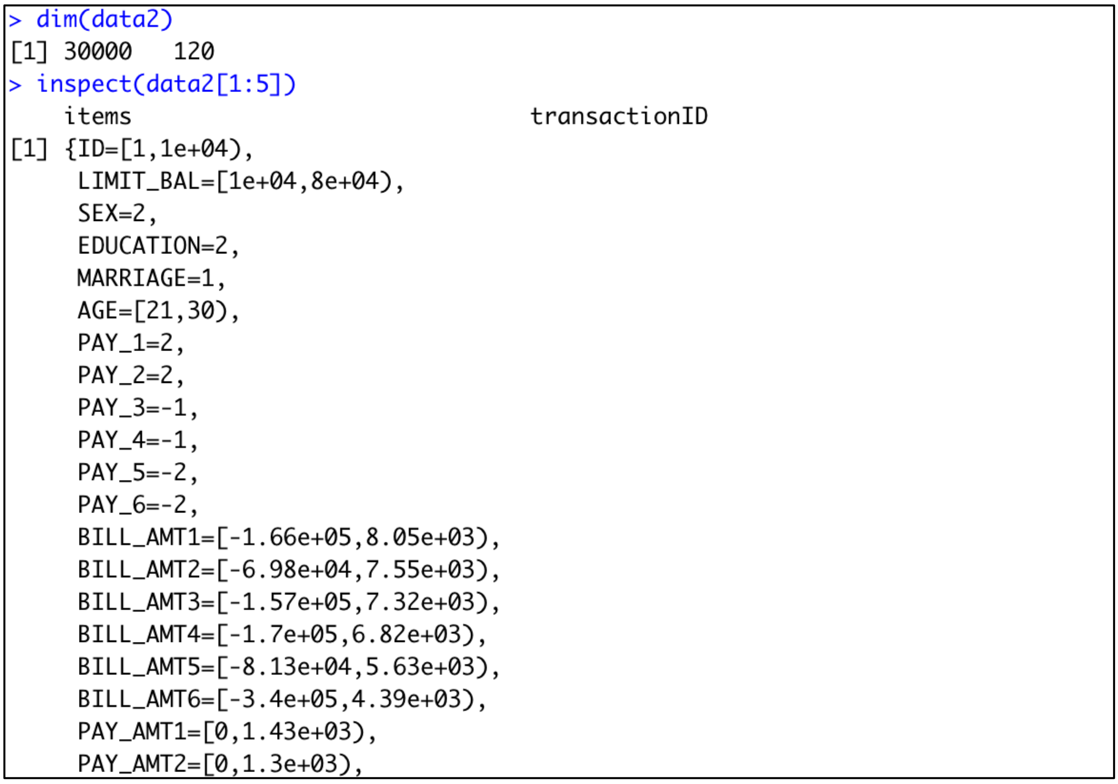
The default graph by gender, marital status and education level

As can be seen from the figure above, women alone account for more breach of contract, but men are more likely to breach of contract than women in horizontal comparison. In terms of education level, college students who perform contracts and breach contracts account for the largest proportion, and customers with postgraduate education are more inclined to repeat contracts. From the perspective of marital status, married people are more likely to breach contracts.

From the perspective of age and credit limit, the age range of default exceeds the performance range, and customers with low credit limits are more likely to default.

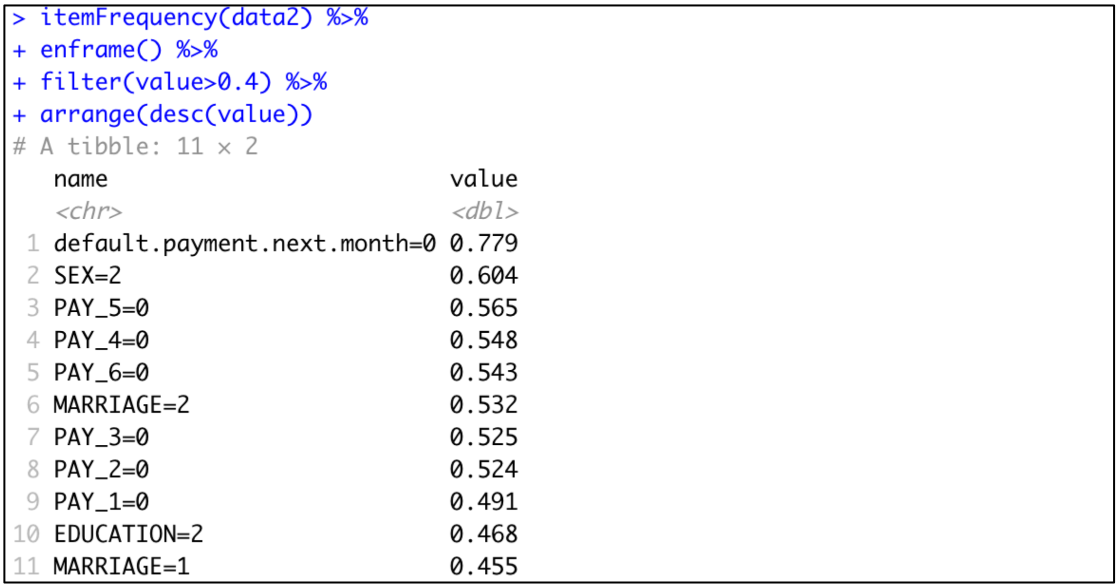
2. Use the association rule algorithm to judge the degree of customer default risk.

(1) Organize and form the data styles required for association rule analysis. The data in the transaction format must be used for association rule.

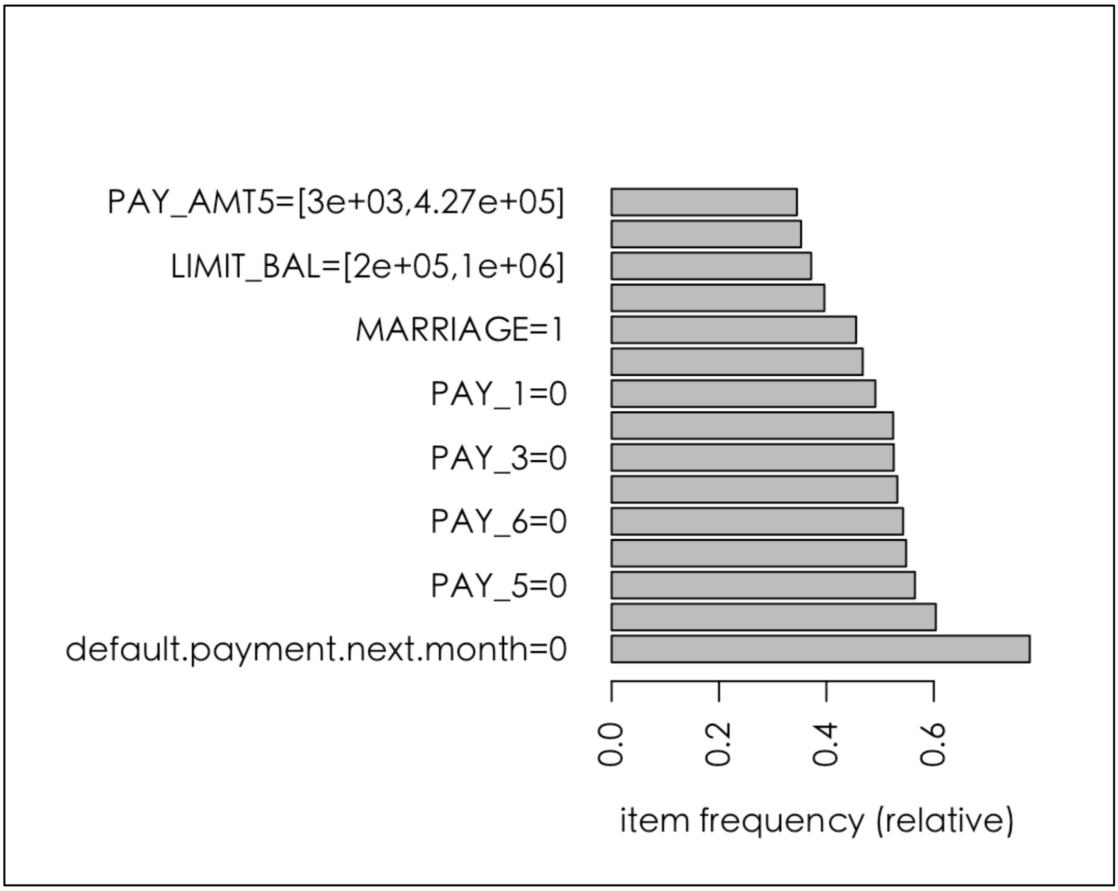


```
> dim(data2)
[1] 30000   120
> inspect(data2[1:5])
    items                            transactionID
[1] {ID=[1,1e+04),
     LIMIT_BAL=[1e+04,8e+04),
     SEX=2,
     EDUCATION=2,
     MARRIAGE=1,
     AGE=[21,30),
     PAY_1=2,
     PAY_2=2,
     PAY_3=-1,
     PAY_4=-1,
     PAY_5=-2,
     PAY_6=-2,
     BILL_AMT1=[-1.66e+05,8.05e+03),
     BILL_AMT2=[-6.98e+04,7.55e+03),
     BILL_AMT3=[-1.57e+05,7.32e+03),
     BILL_AMT4=[-1.7e+05,6.82e+03),
     BILL_AMT5=[-8.13e+04,5.63e+03),
     BILL_AMT6=[-3.4e+05,4.39e+03),
     PAY_AMT1=[0,1.43e+03),
     PAY_AMT2=[0,1.3e+03),
```

(2) Generate the support function and visualize the variable data with the top 15 support;

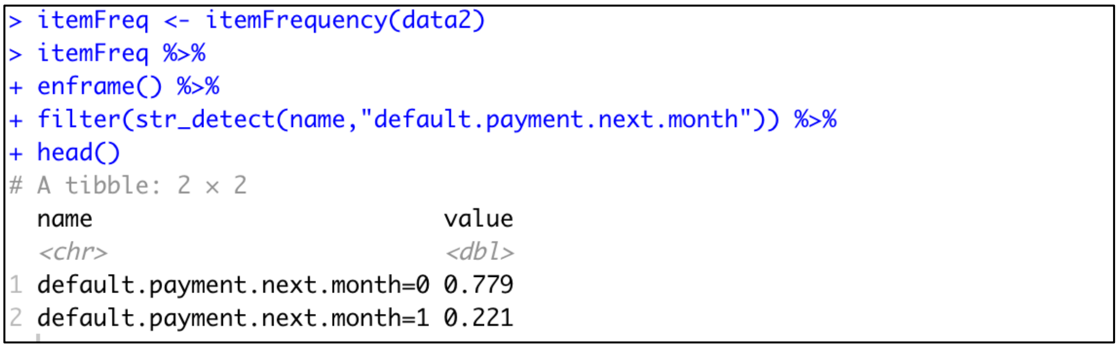


```
> itemFrequency(data2) %>%
+ enframe() %>%
+ filter(value>0.4) %>%
+ arrange(desc(value))
# A tibble: 11 × 2
   name                          value
   <chr>                         <dbl>
 1 default.payment.next.month=0 0.779
 2 SEX=2                         0.604
 3 PAY_5=0                       0.565
 4 PAY_4=0                       0.548
 5 PAY_6=0                       0.543
 6 MARRIAGE=2                    0.532
 7 PAY_3=0                       0.525
 8 PAY_2=0                       0.524
 9 PAY_1=0                       0.491
10 EDUCATION=2                   0.468
11 MARRIAGE=1                    0.455
```



It can be seen that the top support is the default situation in the next month, the monthly repayment situation, the marital status and the credit limit.

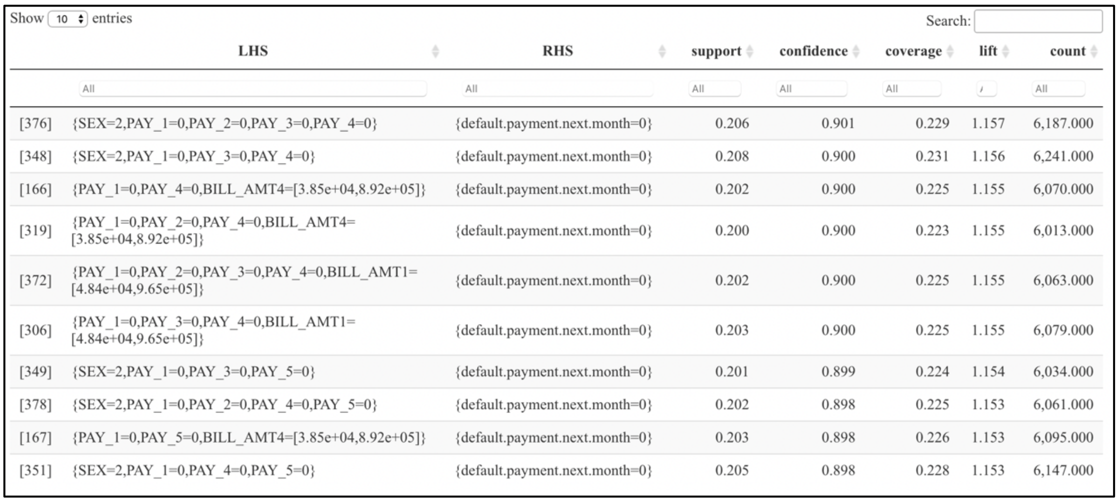
(3) Check the support for default separately; It can be seen that the performance value is 0.779 and the default value is 0.221;



```
> itemFreq <- itemFrequency(data2)
> itemFreq %>%
+ enframe() %>%
+ filter(str_detect(name,"default.payment.next.month")) %>%
+ head()
# A tibble: 2 × 2
  name                         value
  <chr>                        <dbl>
1 default.payment.next.month=0 0.779
2 default.payment.next.month=1 0.221
```

(4) The Apriori algorithm is used for association rule analysis, and the support is set to 0.2 and confidence to 0.8. The appearance parameter is used to control the content of the right item in the rule, and the inspectDT() function is used to check the result. The output result is sorted according to the degree of

improvement.

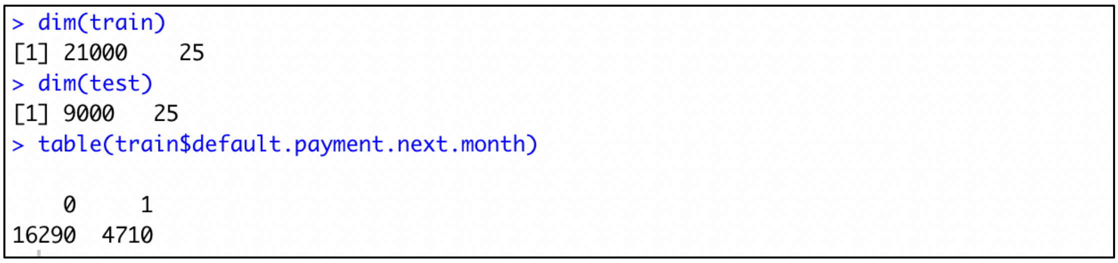


	LHS	RHS	support	confidence	coverage	lift	count
[376]	{SEX=2,PAY_1=0,PAY_2=0,PAY_3=0,PAY_4=0}	{default.payment.next.month=0}	0.206	0.901	0.229	1.157	6,187.000
[348]	{SEX=2,PAY_1=0,PAY_3=0,PAY_4=0}	{default.payment.next.month=0}	0.208	0.900	0.231	1.156	6,241.000
[166]	{PAY_1=0,PAY_4=0,BILL_AMT4=[3.85e+04,8.92e+05]}	{default.payment.next.month=0}	0.202	0.900	0.225	1.155	6,070.000
[319]	{PAY_1=0,PAY_2=0,PAY_4=0,BILL_AMT4=[3.85e+04,8.92e+05]}	{default.payment.next.month=0}	0.200	0.900	0.223	1.155	6,013.000
[372]	{PAY_1=0,PAY_2=0,PAY_3=0,PAY_4=0,BILL_AMT1=[4.84e+04,9.65e+05]}	{default.payment.next.month=0}	0.202	0.900	0.225	1.155	6,063.000
[306]	{PAY_1=0,PAY_3=0,PAY_4=0,BILL_AMT1=[4.84e+04,9.65e+05]}	{default.payment.next.month=0}	0.203	0.900	0.225	1.155	6,079.000
[349]	{SEX=2,PAY_1=0,PAY_3=0,PAY_5=0}	{default.payment.next.month=0}	0.201	0.899	0.224	1.154	6,034.000
[378]	{SEX=2,PAY_1=0,PAY_2=0,PAY_4=0,PAY_5=0}	{default.payment.next.month=0}	0.202	0.898	0.225	1.153	6,061.000
[167]	{PAY_1=0,PAY_5=0,BILL_AMT4=[3.85e+04,8.92e+05]}	{default.payment.next.month=0}	0.203	0.898	0.226	1.153	6,095.000
[351]	{SEX=2,PAY_1=0,PAY_4=0,PAY_5=0}	{default.payment.next.month=0}	0.205	0.898	0.228	1.153	6,147.000

From the first result for example, if a customer's gender is female, and from January to April are timely repayment, then she will perform the contract in the next month, the support of this rule is 0.206, confidence is 0.901, indicating that nearly 10% confidence level (non-default probability is 90%) can cover about 20% of customers. Compared with the male who did not repay in time from January to April, her probability of not defaulting in the next month increased by 1.157 times. The other rules are similar, so I won't elaborate on them.

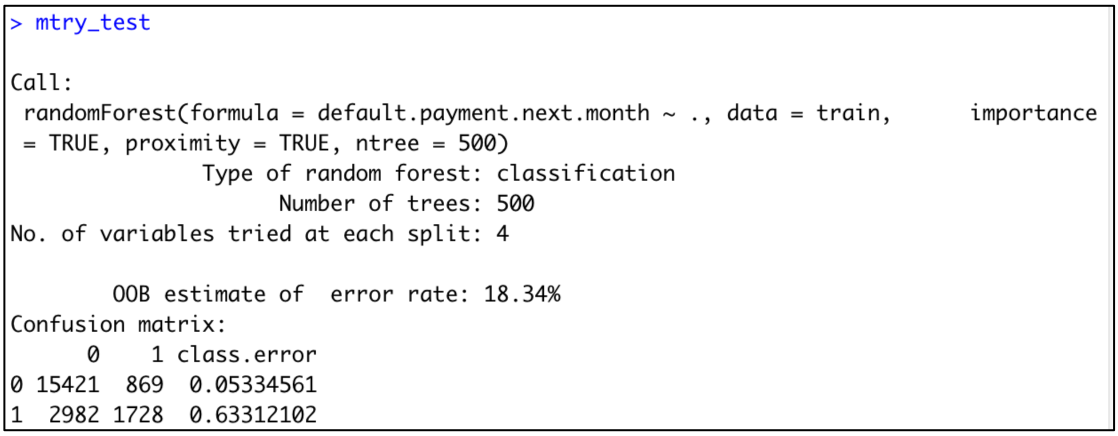
3. Use random forest algorithm to model and predict defaulting customers.

(1) Import random forest package to 'default. Payment. Next month' as a target variable, and the data is divided into training set and test set, 70% of the original data as the training set, and the rest for the test set, to view the situation. 21000 can be seen in the training set data, test sets, and 9000 training. The default payment. Next, the month of the variable code 0, 16290, 1, 4710;

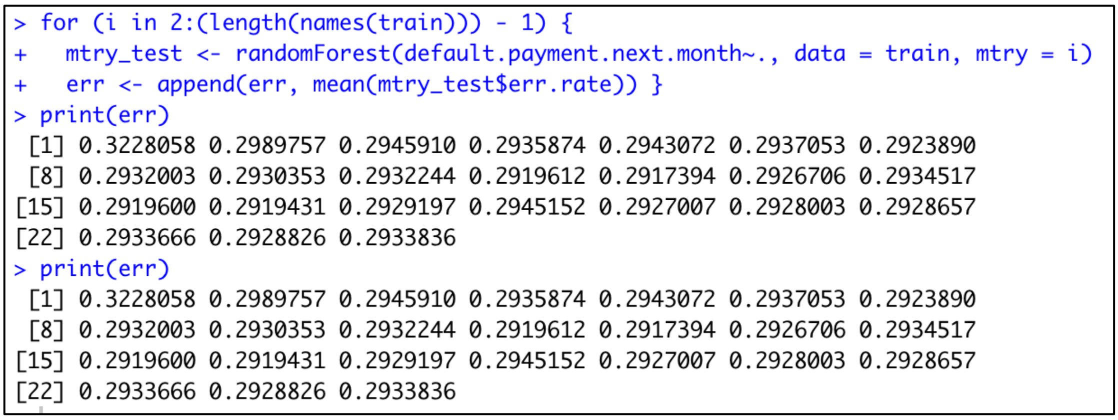


```
> dim(train)
[1] 21000    25
> dim(test)
[1] 9000   25
> table(train$default.payment.next.month)

    0     1 
16290  4710 
```

(2) The randomForest() function is used to build the model, and the decision tree and error probability distribution diagram are viewed to determine the number of decision trees. It can be seen that when the decision tree is 500, the OOB error is 18.34%, and it can be seen from the figure that with the increase of decision tree types, the error rate shows a downward trend, reaching about 1000 error rate change is not obvious, so choose ntree=1000 as the final classifier;

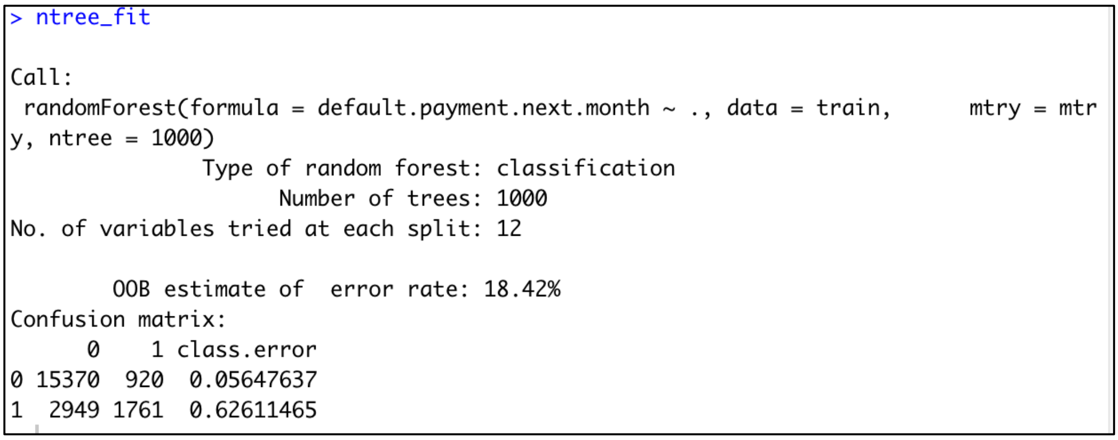


```
> mtry_test

Call:
 randomForest(formula = default.payment.next.month ~ ., data = train,      importance
 = TRUE, proximity = TRUE, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 18.34%
Confusion matrix:
      0    1 class.error
0 15421  869  0.05334561
1  2982 1728  0.63312102
```

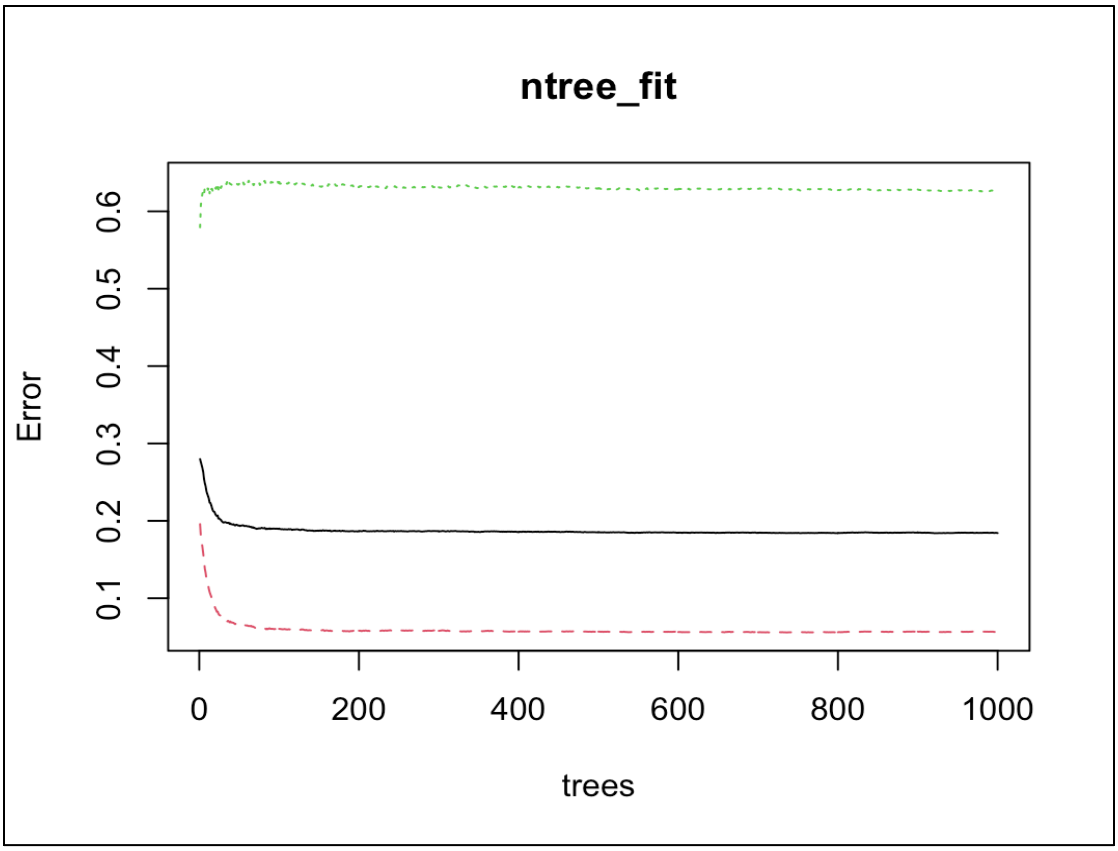


```
> for (i in 2:(length(names(train))) - 1) {
+   mtry_test <- randomForest(default.payment.next.month~., data = train, mtry = i)
+   err <- append(err, mean(mtry_test$err.rate)) }
> print(err)
 [1] 0.3228058 0.2989757 0.2945910 0.2935874 0.2943072 0.2937053 0.2923890
 [8] 0.2932003 0.2930353 0.2932244 0.2919612 0.2917394 0.2926706 0.2934517
[15] 0.2919600 0.2919431 0.2929197 0.2945152 0.2927007 0.2928003 0.2928657
[22] 0.2933666 0.2928826 0.2933836
> print(err)
 [1] 0.3228058 0.2989757 0.2945910 0.2935874 0.2943072 0.2937053 0.2923890
 [8] 0.2932003 0.2930353 0.2932244 0.2919612 0.2917394 0.2926706 0.2934517
[15] 0.2919600 0.2919431 0.2929197 0.2945152 0.2927007 0.2928003 0.2928657
[22] 0.2933666 0.2928826 0.2933836
```

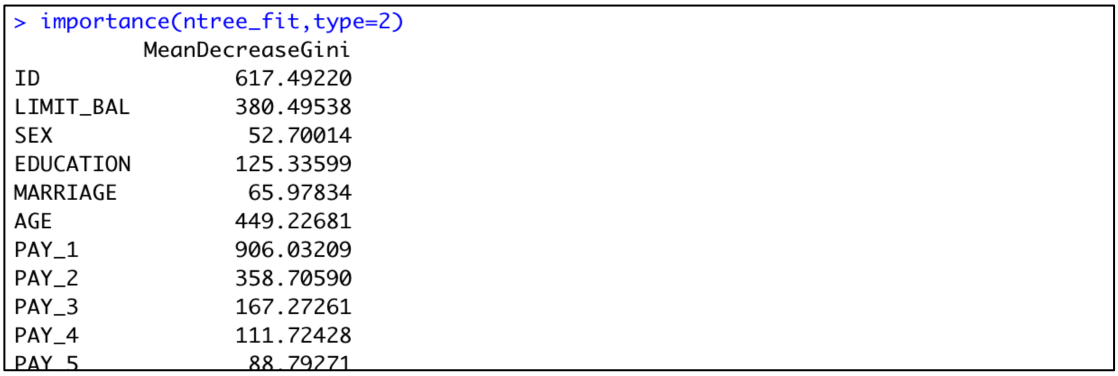
(3) The final classifier is obtained, and the model is built. As can be seen from the results, the second and third categories still have errors and will be misjudged;



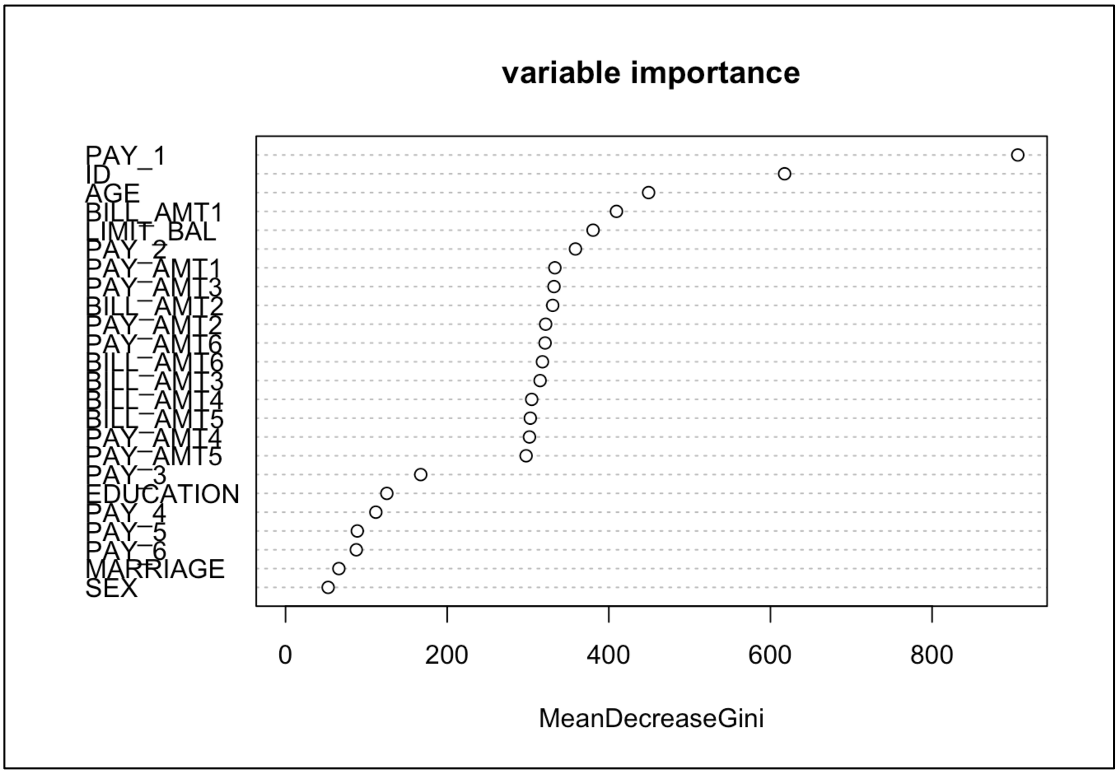
```
> ntree_fit

Call:
 randomForest(formula = default.payment.next.month ~ ., data = train,      mtry = mtr
y, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 12

        OOB estimate of  error rate: 18.42%
Confusion matrix:
      0    1 class.error
0 15370  920  0.05647637
1  2949 1761  0.62611465
```



(4) The importance of indicators is visualized utilizing graphs; It can be seen that the repayment in January, age, and credit limit are very important to the prediction result;

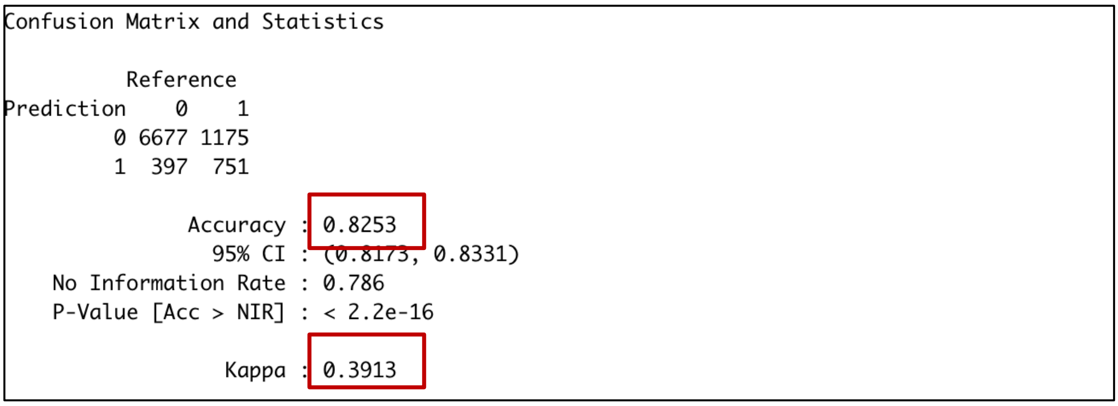


```
> importance(ntree_fit,type=2)
          MeanDecreaseGini
ID               617.49220
LIMIT_BAL        380.49538
SEX               52.70014
EDUCATION        125.33599
MARRIAGE          65.97834
AGE              449.22681
PAY_1            906.03209
PAY_2            358.70590
PAY_3            167.27261
PAY_4            111.72428
PAY_5             88.79271
```



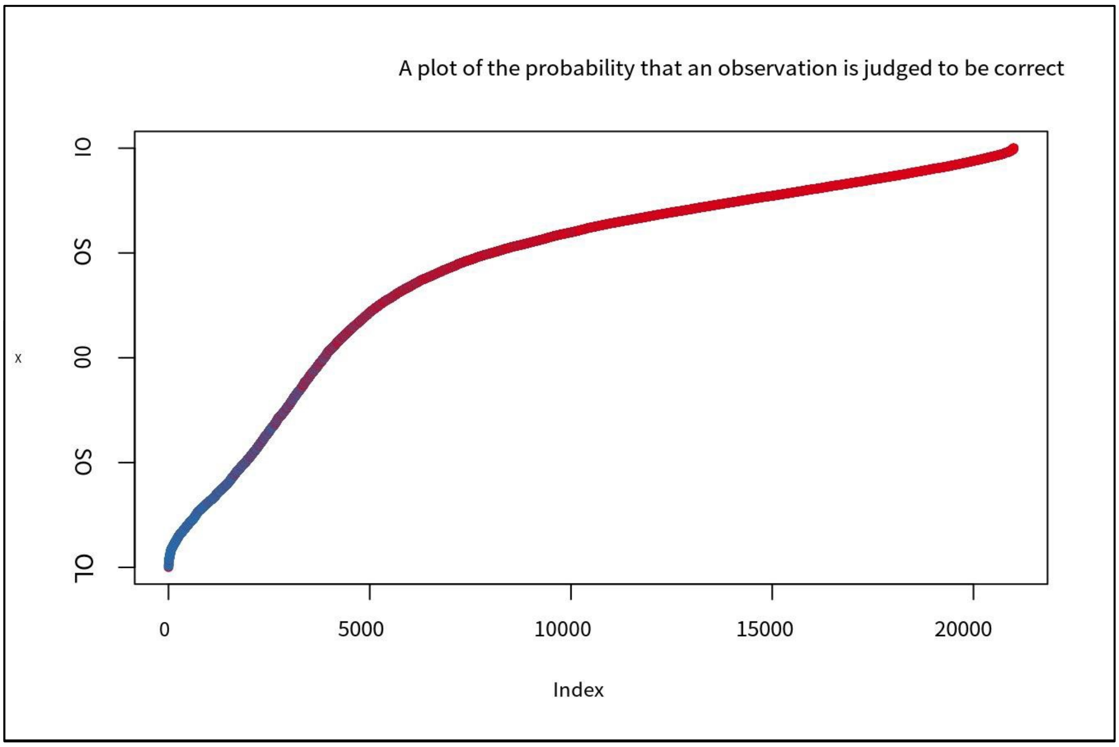
4. Confusion matrix and ROC curve evaluate the prediction effect

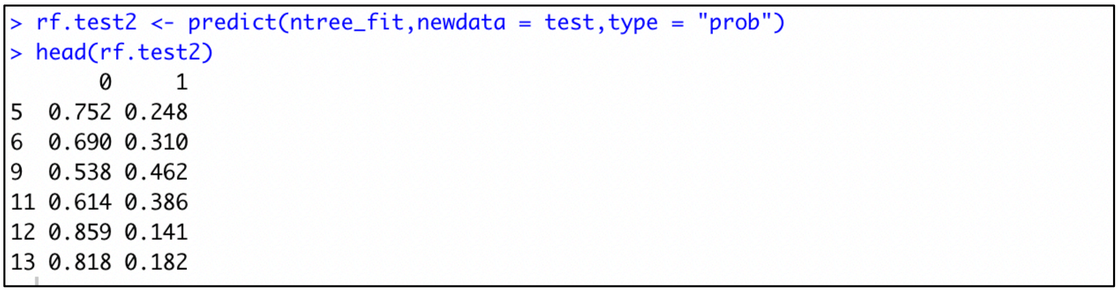
(1) The data of the test set were predicted, and the confusion matrix was output, and the probability of the observed value being accurately observed was drawn to evaluate the prediction effect; It can be seen that the accuracy rate of this model reached 82.53%, kappa value was 0.3913, sensitivity was 94.39%, and balance accuracy was 66.69%. Most of the samples had a very high accuracy rate for classification. Generally, the effect was good.



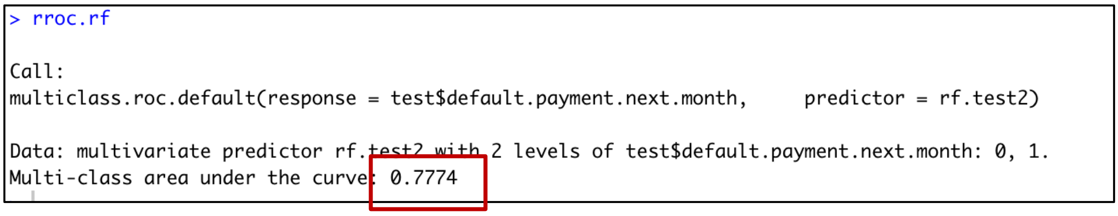
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6677 1175
         1  397  751

               Accuracy : 0.8253
                 95% CI : (0.8173, 0.8331)
    No Information Rate : 0.786
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3913
```

(2) The binary classification roc() function was used to judge the quality of the model and ROC curve was drawn. It could be seen that the value of AUC was 77.74%, close to 80%, indicating that the model had good effect.

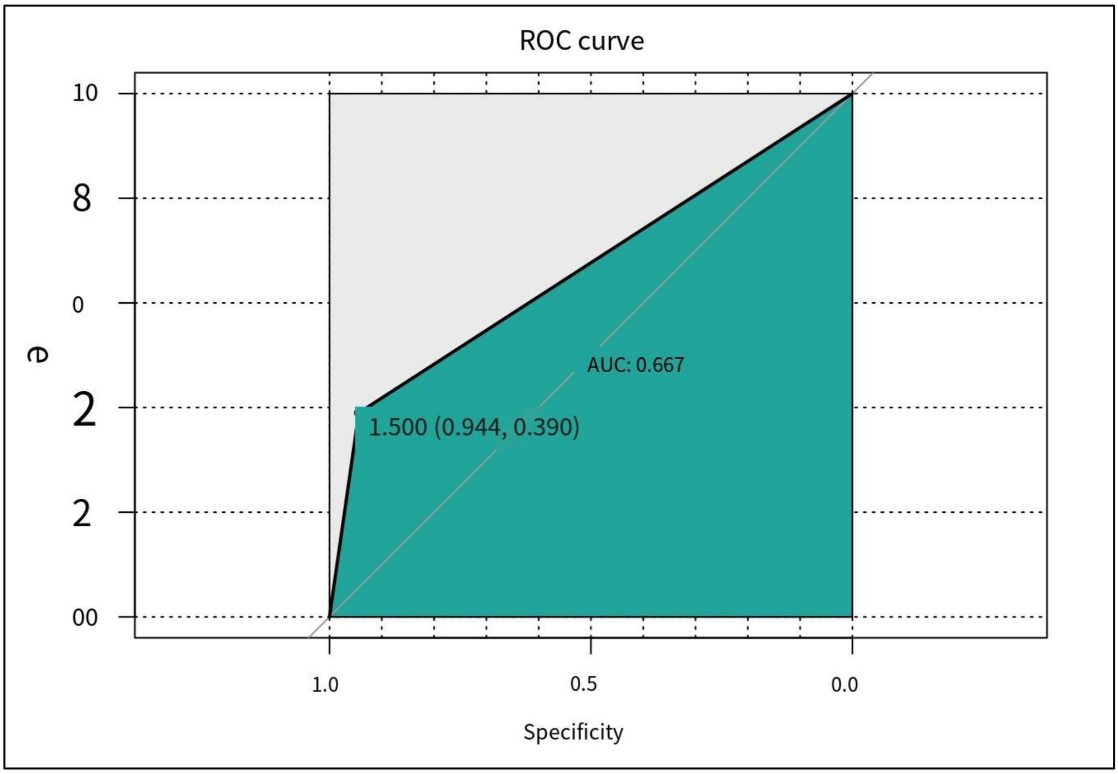

A plot of the probability that an observation is judged to be correct



```
> rf.test2 <- predict(ntree_fit,newdata = test,type = "prob")
> head(rf.test2)
       0     1
5  0.752 0.248
6  0.690 0.310
9  0.538 0.462
11 0.614 0.386
12 0.859 0.141
13 0.818 0.182
```



```
> rroc.rf

Call:
multiclass.roc.default(response = test$default.payment.next.month,     predictor = rf.test2)

Data: multivariate predictor rf.test2 with 2 levels of test$default.payment.next.month: 0, 1.
Multi-class area under the curve: 0.7774
```



IV. Experimental conclusions

From the above experiments, it can be seen that the association rule algorithm can be used to analyze the female customers in the credit card customer data set who pay monthly are the least likely to default (in order of improvement degree). The support degree of this rule is 0.206, the confidence degree is 0.901, which can cover about 20% of customers, and the probability of not defaulting is 90.1%. While male customers who do not pay back every month are more likely to default; By using the random effect model to forecast the customer default modeling in the next month, it can be seen that whether monthly repayment, age and credit limit are the most important to the prediction. The accuracy of the prediction data by using the random forest model reaches 82.53%, the sensitivity is 94.39%, and the robustness is 66.695. In addition, the roc() function was used to evaluate the effect of probabilistic (continuous) prediction data. It can be seen that the AUC value reached 77.74% and the numerical (discrete) value was 66.7%. It can be judged that the model has a good prediction effect on customer default risk, but there are still some errors taken into account, such as the overfitting phenomenon that may occur due to excessive noise. Or the attributes with more level divisions affect the prediction effect of random forest and need to be optimized.