Titanic

Han Wang

2017年8月16日

1. 数据结构

```
train<-read.csv("D:/LearningR/Titanic/train.csv")
test<-read.csv("D:/LearningR/Titanic/test.csv")
model<-read.csv("D:/LearningR/Titanic/gender_submission.csv")
test$Survived<-model$Survived#[match(test$PassengerID, Real$PassengerID)]
summary(test)
```

```
##   PassengerId         Pclass
##  Min.   : 892.0   Min.   :1.000
##  1st Qu.: 996.2   1st Qu.:1.000
##  Median :1100.5   Median :3.000
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##                                          Name         Sex
##  Abbott, Master. Eugene Joseph            :  1   female:152
##  Abelseth, Miss. Karen Marie              :  1   male  :266
##  Abelseth, Mr. Olaus Jorgensen            :  1
##  Abrahamsson, Mr. Abraham August Johannes :  1
##  Abrahim, Mrs. Joseph (Sophie Halaut Easu):  1
##  Aks, Master. Philip Frank                :  1
##  (Other)                                  :412
##       Age            SibSp            Parch             Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   PC 17608:  5
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   113503  :  4
##  Median :27.00   Median :0.0000   Median :0.0000   CA. 2343:  4
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923   16966   :  3
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000   220845  :  3
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000   347077  :  3
##  NA's   :86                                        (Other) :396
##       Fare                   Cabin     Embarked    Survived
##  Min.   :  0.000                :327   C:102    Min.   :0.0000
##  1st Qu.:  7.896   B57 B59 B63 B66:  3   Q: 46    1st Qu.:0.0000
##  Median : 14.454   A34           :  2   S:270    Median :0.0000
##  Mean   : 35.627   B45           :  2            Mean   :0.3636
##  3rd Qu.: 31.500   C101          :  2            3rd Qu.:1.0000
##  Max.   :512.329   C116          :  2            Max.   :1.0000
##  NA's   :1         (Other)       : 80
```

```
head(test)
```

```
##   PassengerId Pclass                                          Name    Sex
## 1         892      3                              Kelly, Mr. James   male
## 2         893      3              Wilkes, Mrs. James (Ellen Needs) female
## 3         894      2                     Myles, Mr. Thomas Francis   male
## 4         895      3                              Wirz, Mr. Albert   male
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female
## 6         897      3                    Svensson, Mr. Johan Cervin   male
##    Age SibSp Parch  Ticket    Fare Cabin Embarked Survived
## 1 34.5     0     0  330911  7.8292              Q        0
## 2 47.0     1     0  363272  7.0000              S        1
## 3 62.0     0     0  240276  9.6875              Q        0
## 4 27.0     0     0  315154  8.6625              S        0
## 5 22.0     1     1 3101298 12.2875              S        1
## 6 14.0     0     0    7538  9.2250              S        0
```

```
test$Survived<-model$Survived#[match(test$PassengerID, Real$PassengerID)]
full  <- rbind(train, test) # bind training & test data

str(full)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417
 581 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 929 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133
 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
full$Title <- gsub('(.*, )|(\\..*)', '', full$Name)#去除“, ”之前的任意字符以及“."后的任意字符
full$Survived<-as.factor(full$Survived)
full$Pclass<-as.factor(full$Pclass)
full$Title<-as.factor(full$Title)
head(full)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                             Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                              Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                            Allen, Mr. William Henry   male  35     0
## 6                                    Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked Title
## 1     0        A/5 21171  7.2500              S    Mr
## 2     0         PC 17599 71.2833   C85        C   Mrs
## 3     0 STON/O2. 3101282  7.9250              S  Miss
## 4     0           113803 53.1000  C123        S   Mrs
## 5     0           373450  8.0500              S    Mr
## 6     0           330877  8.4583              Q    Mr
```

```
features_1<-full[,c(3,5,6,7,8,10,11,12)]
#target_train<-train[,c(2)]
str(features_1)
```

```
## 'data.frame':    1309 obs. of  8 variables:
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin   : Factor w/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
features_2<-full[,c(3,5,6,7,8,10,11,12,13)]
str(features_2)
```

```
## 'data.frame':    1309 obs. of  9 variables:
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin   : Factor w/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
##  $ Title   : Factor w/ 18 levels "Capt","Col","Don",..: 13 14 10 14 13 13 13 9 14 14 ...
```

2. 描述统计

```
#性别
train<-full[c(1:891),]
table(train$Survived,train$Sex)
```

```
##
##     female male
##   0     81  468
##   1    233  109
```

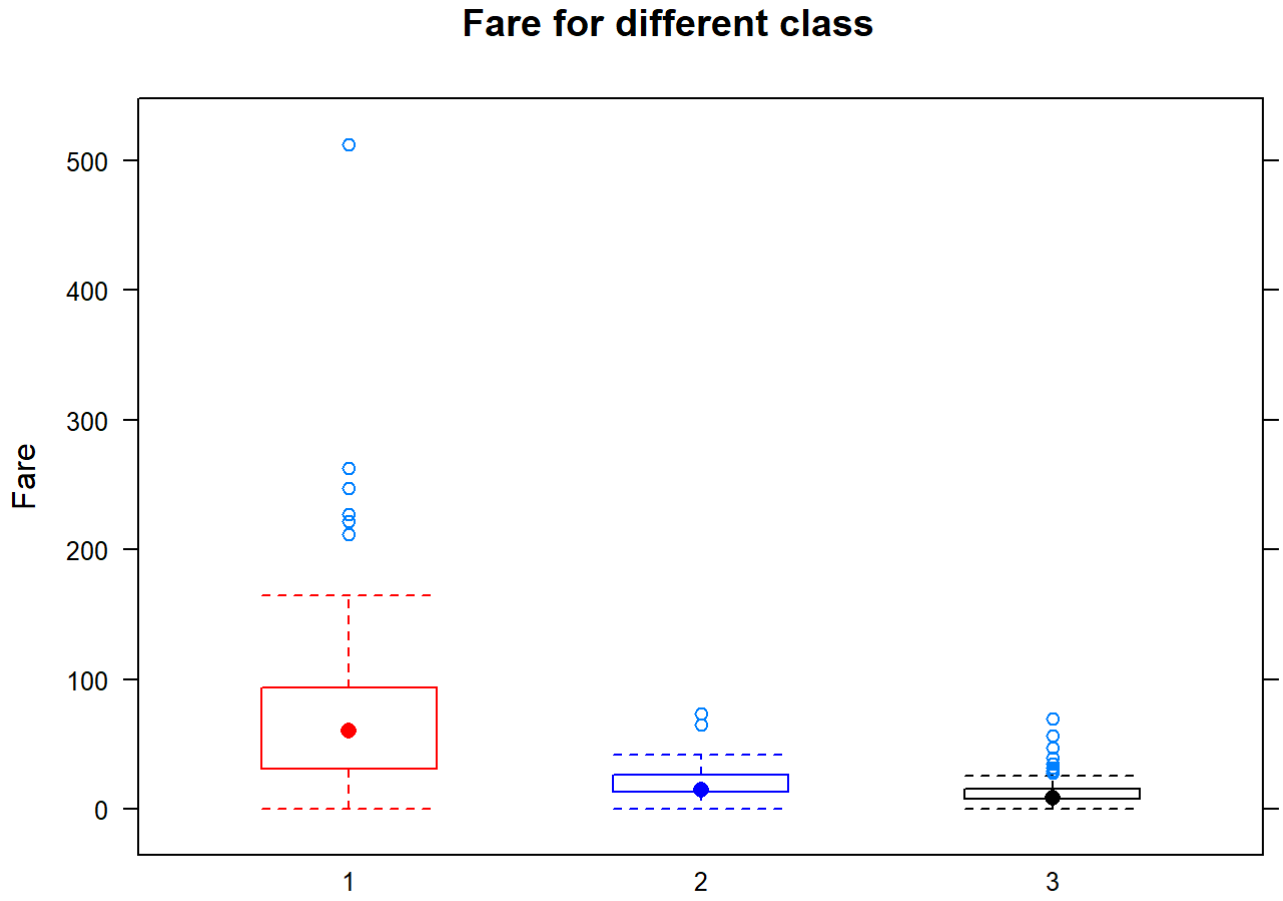
```
#船舱
table(train$Survived,train$Pclass)
```

```
##
##       1   2   3
##   0  80  97 372
##   1 136  87 119
```

```
#船舱与称呼的联系
table(train$Title,train$Pclass)
```

```
##
##                  1   2   3
##   Capt           1   0   0
##   Col            2   0   0
##   Don            1   0   0
##   Dona           0   0   0
##   Dr             5   2   0
##   Jonkheer       1   0   0
##   Lady           1   0   0
##   Major          2   0   0
##   Master         3   9  28
##   Miss          46  34 102
##   Mlle           2   0   0
##   Mme            1   0   0
##   Mr           107  91 319
##   Mrs           42  41  42
##   Ms             0   1   0
##   Rev            0   6   0
##   Sir            1   0   0
##   the Countess   1   0   0
```

```
bwplot( Fare ~ Pclass, data=train,main="Fare for different class",par.settings = list(box.umbrella=list(col= c("red", "blue","black")), box.dot=list(col= c("red", "blue","black")), box.rectangle = list(col= c("red", "blue","black"))))
```



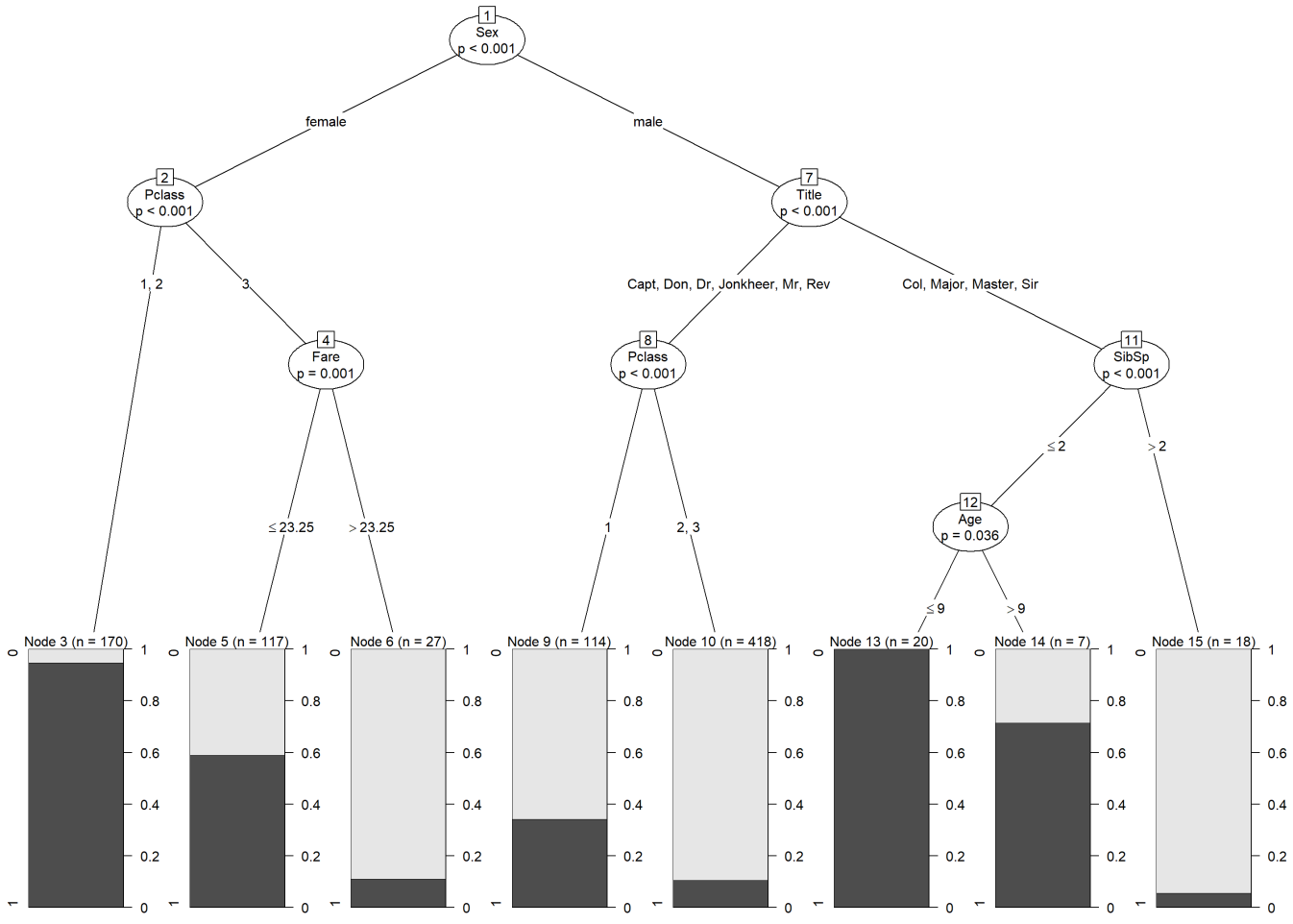
粗略来看女性存活率高于男性，头等舱存活率高于二等舱远高于三等舱 地位高的人大部分在头等舱，例如上校、伯爵夫人、贵族等；神父大多在二等舱 头等舱票价是三等舱票价的5倍以上

3. 条件推理树(conditional inferences tree)

```
CItree<- ctree(Survived~Pclass+Sex+Title+Age+SibSp+Parch+Fare+Embarked,data=train)
CItree
```

```
##
## Model formula:
## Survived ~ Pclass + Sex + Title + Age + SibSp + Parch + Fare +
##     Embarked
##
## Fitted party:
## [1] root
## |   [2] Sex in female
## |   |   [3] Pclass in 1, 2: 1 (n = 170, err = 5.3%)
## |   |   [4] Pclass in 3
## |   |   |   [5] Fare <= 23.25: 1 (n = 117, err = 41.0%)
## |   |   |   [6] Fare > 23.25: 0 (n = 27, err = 11.1%)
## |   [7] Sex in male
## |   |   [8] Title in Capt, Don, Dr, Jonkheer, Mr, Rev
## |   |   |   [9] Pclass in 1: 0 (n = 114, err = 34.2%)
## |   |   |   [10] Pclass in 2, 3: 0 (n = 418, err = 10.5%)
## |   |   [11] Title in Col, Major, Master, Sir
## |   |   |   [12] SibSp <= 2
## |   |   |   |   [13] Age <= 9: 1 (n = 20, err = 0.0%)
## |   |   |   |   [14] Age > 9: 1 (n = 7, err = 28.6%)
## |   |   |   [15] SibSp > 2: 0 (n = 18, err = 5.6%)
##
## Number of inner nodes:    7
## Number of terminal nodes: 8
```

```
plot(CItree)
```

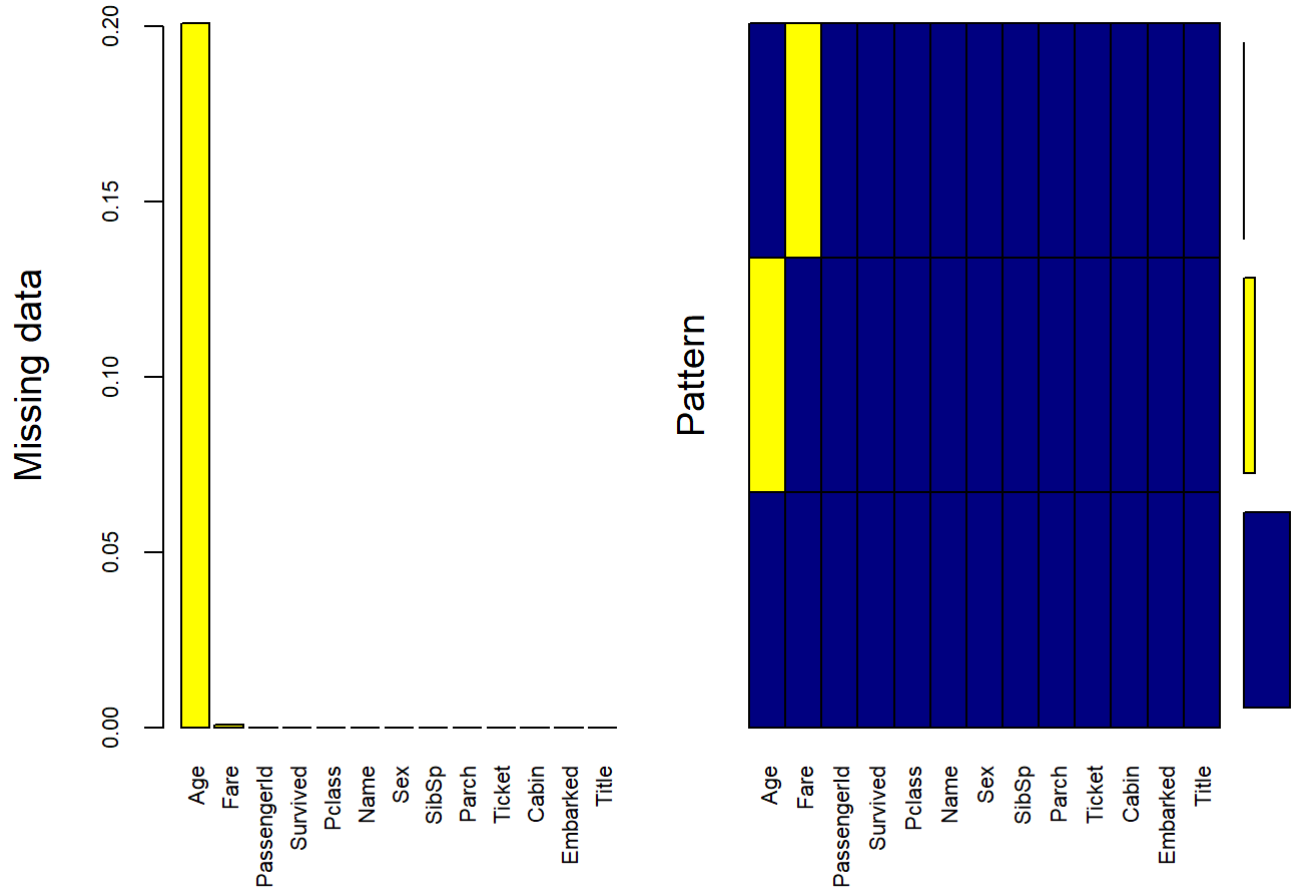


结果与发现： 一、二等舱女性存活率是三等舱女性的两倍多； 男性中二三等舱男性以及船舱上有较多家庭成员数（2+）的存活率最低。

4.数据预处理

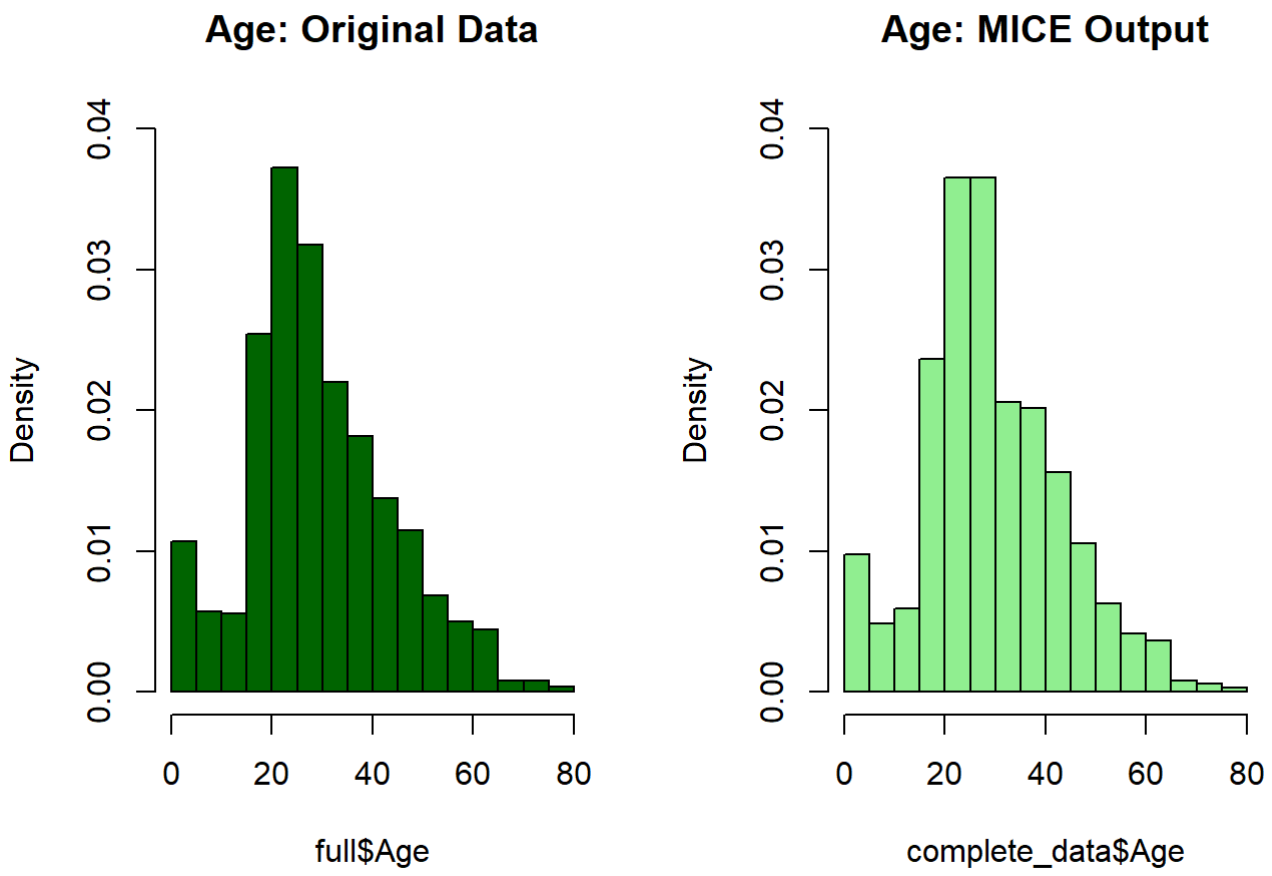
```
#缺失值模式
aggr(full, col=c('navyblue','yellow'),numbers=TRUE, sortVars=TRUE,labels=names(full), cex.axis=.7,
gap=3, ylab=c("Missing data","Pattern"))
```

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
##
##  Variables sorted by number of missings:
##     Variable        Count
##          Age 0.2009167303
##         Fare 0.0007639419
##  PassengerId 0.0000000000
##     Survived 0.0000000000
##       Pclass 0.0000000000
##         Name 0.0000000000
##          Sex 0.0000000000
##        SibSp 0.0000000000
##        Parch 0.0000000000
##       Ticket 0.0000000000
##        Cabin 0.0000000000
##     Embarked 0.0000000000
##        Title 0.0000000000
```

```
#20% 年龄缺失 以及少量Fare缺失 采用multiple imputation
#impute_dataset<-full[,c(3,5,6,7,8,10,11,12,13)]#用除Name,Ticket 以外的变量建模-
set.seed(123)
#imputed<-mice(impute_dataset,m=5,maxit = 50,seed=500)
#summary(imputed)
#mice_output <- complete(imputed,3)#随机选第三个
#write.csv(mice_output,"D:/LearningR/Titanic/new_train.csv")
complete_data<-read.csv("D:/LearningR/Titanic/new_train.csv")#complete data
#plot the distribution
par(mfrow=c(1,2))
hist(full$Age, freq=F, main='Age: Original Data',
  col='darkgreen', ylim=c(0,0.04))
hist(complete_data$Age, freq=F, main='Age: MICE Output',
  col='lightgreen', ylim=c(0,0.04))
```

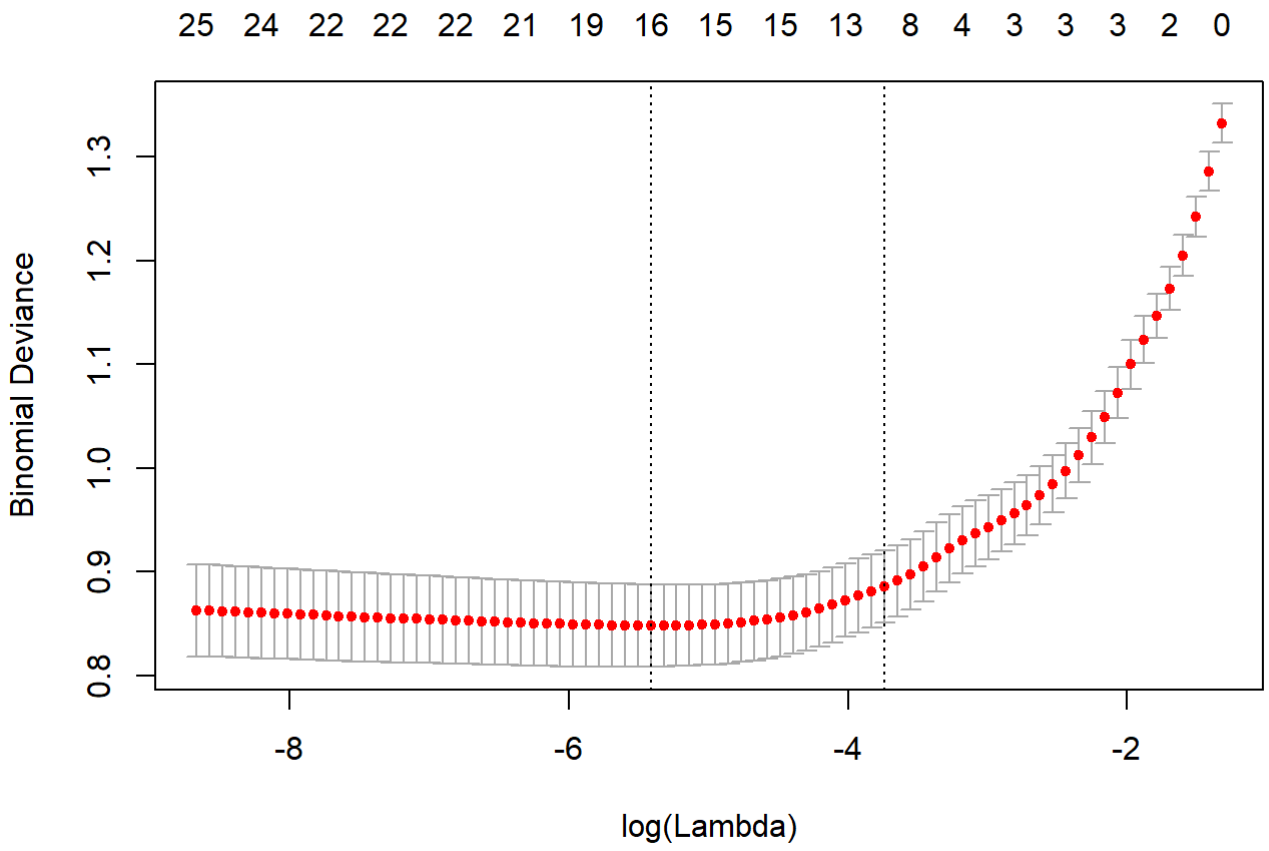


5. 数据分区

```
set.seed(123)
complete_data$Survived<-full$Survived
new_train<-complete_data[c(1:891),]
new_test<-complete_data[c(892:1309),]
#不显示warning
options(warn=-1)
```

6. LASSO 回归-特征选择

```
#建变量矩阵
xfactors<-model.matrix(~Pclass+Sex+Title+Age+SibSp+Parch+Fare+Embarked,data=new_train)
TrainY<-new_train$Survived
#交叉验证
set.seed(123)
CV_LASSO<-cv.glmnet(x=xfactors,y=TrainY,family='binomial',alpha=1)
plot(CV_LASSO)
```




```
#选使standard error 最小的lambda
fit_LASSO<-glmnet(x=xfactors,y=TrainY,family='binomial',alpha=1,lambda = CV_LASSO$lambda.1se)
#结果
coef(fit_LASSO)
```

```
## 28 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)    2.74313664
## (Intercept)    .
## Pclass        -0.73500959
## Sexmale       -1.73640795
## TitleCol       .
## TitleCountess  .
## TitleDon       .
## TitleDona      .
## TitleDr        .
## TitleJonkheer  .
## TitleLady      .
## TitleMajor     .
## TitleMaster    1.27834123
## TitleMiss      .
## TitleMlle      .
## TitleMme       .
## TitleMr       -0.80558088
## TitleMrs       .
## TitleMs        .
## TitleRev      -0.34645539
## TitleSir       .
## Age            .
## SibSp         -0.19878153
## Parch          .
## Fare           .
## EmbarkedC      0.11109328
## EmbarkedQ      .
## EmbarkedS     -0.07196896
```

Lasso结果： 同进同出原则，去掉Age,Parch,Fare

7. 模型预测

a. 逻辑回归

```
#交叉验证参数
train_control<- trainControl(method="cv", number=10)
#模型
set.seed(123)
fit_LR<- train(Survived~Pclass+Sex+Title+Age+SibSp+Parch+Fare+Embarked,data=new_train,  trControl=trai
n_control, method="glm", family=binomial(),metric="Accuracy")



#全部变量放进去
pred <- predict(fit_LR, newdata=new_test)


#使用LASSO回归结果
set.seed(123)
fit_LR_2<- train(Survived~Pclass+Sex+Title+SibSp+Embarked,data=new_train,  trControl=train_control, me
thod="glm", family=binomial(),metric="Accuracy")

pred_2<- predict(fit_LR_2, newdata=new_test)
```

b. 随机森林

```
set.seed(123)
#全放
rf.model_1<-train(Survived~Pclass+Sex+Title+Age+SibSp+Parch+Fare+Embarked,data=new_train,method="rf",
trControl=train_control,metric="Accuracy")
```

```
## Loading required package: randomForest
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:Hmisc':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
summary(rf.model_1)
```

```
##                 Length Class      Mode
## call               4   -none-     call
## type               1   -none-     character
## predicted        891   factor     numeric
## err.rate        1500   -none-     numeric
## confusion          6   -none-     numeric
## votes           1782   matrix     numeric
## oob.times        891   -none-     numeric
## classes            2   -none-     character
## importance        26   -none-     numeric
## importanceSD       0   -none-     NULL
## localImportance    0   -none-     NULL
## proximity          0   -none-     NULL
## ntree              1   -none-     numeric
## mtry               1   -none-     numeric
## forest            14   -none-     list
## y                891   factor     numeric
## test               0   -none-     NULL
## inbag              0   -none-     NULL
## xNames            26   -none-     character
## problemType        1   -none-     character
## tuneValue          1   data.frame list
## obsLevels          2   -none-     character
## param              0   -none-     list
```

```
pred.result<-predict(rf.model_1,new_test)

#LASSO 回归结果
set.seed(123)
rf.model_LASSO<-train(Survived~Pclass+Sex+Title+SibSp+Embarked,data=new_train,method="rf", trControl=t
rain_control,metric="Accuracy")
summary(rf.model_LASSO)
```

```
##                 Length Class      Mode
## call               4   -none-     call
## type               1   -none-     character
## predicted        891   factor     numeric
## err.rate        1500   -none-     numeric
## confusion          6   -none-     numeric
## votes           1782   matrix     numeric
## oob.times        891   -none-     numeric
## classes            2   -none-     character
## importance        23   -none-     numeric
## importanceSD       0   -none-     NULL
## localImportance    0   -none-     NULL
## proximity          0   -none-     NULL
## ntree              1   -none-     numeric
## mtry               1   -none-     numeric
## forest            14   -none-     list
## y                891   factor     numeric
## test               0   -none-     NULL
## inbag              0   -none-     NULL
## xNames            23   -none-     character
## problemType        1   -none-     character
## tuneValue          1   data.frame list
## obsLevels          2   -none-     character
## param              0   -none-     list
```

```
pred.result_LASSO<-predict(rf.model_LASSO,new_test)

#write.csv(pred.result,"D:/LearningR/Titanic/perdiction_RF_1.csv")
#write.csv(pred.result_LASSO,"D:/LearningR/Titanic/perdiction_RF_LASSO.csv")
```

C.支持向量机

```
#fit
trctrl <- trainControl(method = "cv", number = 10)
set.seed(123)

#全放
svm_Radial_1 <- train(Survived~Pclass+Sex+Title+Age+SibSp+Parch+Fare+Embarked,data=new_train,method =
"svmRadial", trControl=trctrl, preProcess = c("center", "scale"), tuneLength = 9,metric="Accuracy")
```

```
## Loading required package: kernlab
```

```
##
## Attaching package: 'kernlab'
```

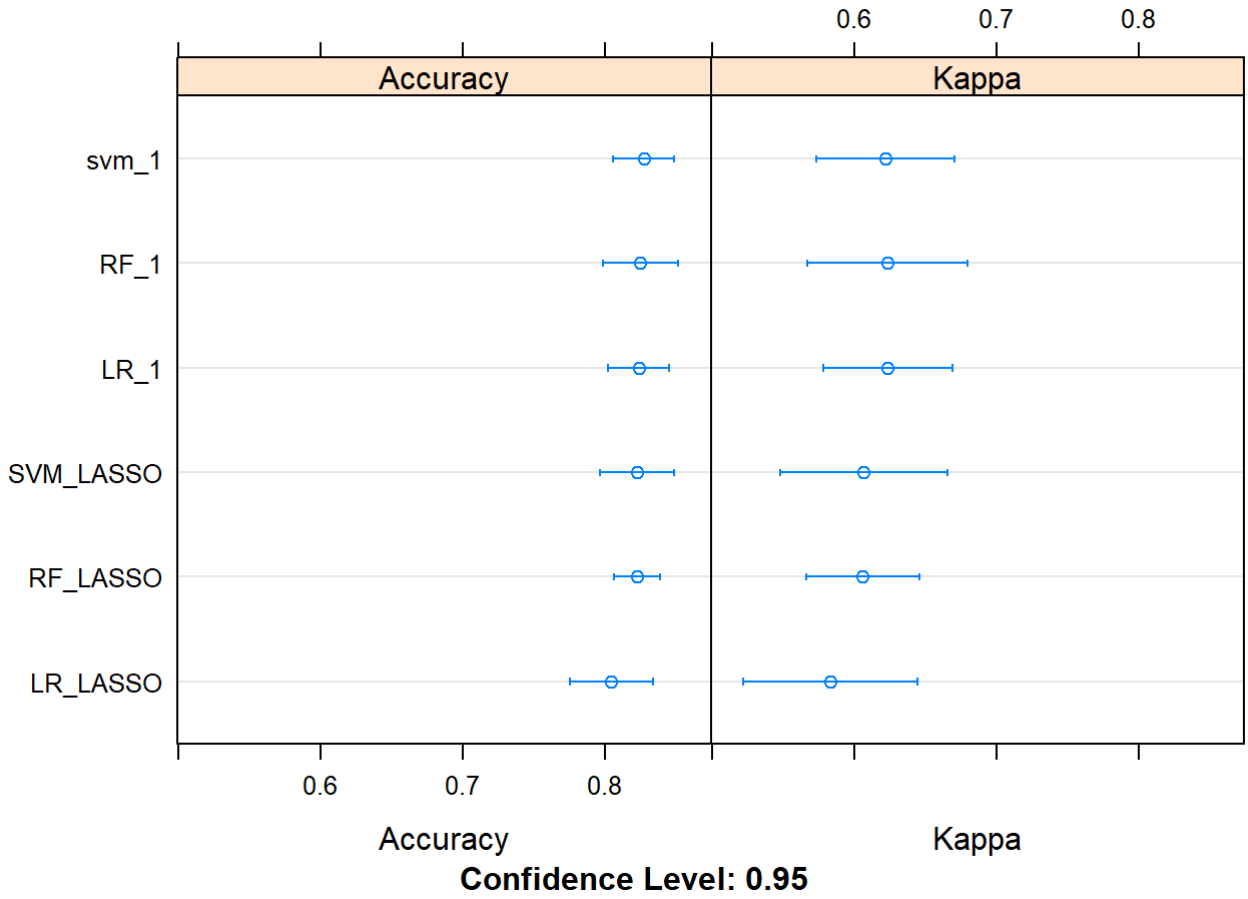
```
## The following object is masked from 'package:modeltools':
##
##     prior
```

```
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
# 预测
pred_SVM_1<-predict(svm_Radial_1,newdata=new_test)
#write.csv(pred_SVM_1,"D:/LearningR/Titanic/perdiction_SVM_1.csv")

#Lasso 结果

set.seed(123)
svm_Radial_LASSO <- train(Survived~Pclass+Sex+Title+SibSp+Embarked,data=new_train,method =
"svmRadial", trControl=trctrl, preProcess = c("center", "scale"), tuneLength = 9,metric="Accuracy")


#预测
pred_SVM_LASSO<-predict(svm_Radial_LASSO,newdata=new_test)


#SVM,随机森林和逻辑回归交叉验证比较
CV<-resamples(list(LR_1=fit_LR,LR_LASSO=fit_LR_2,RF_1=rf.model_1,RF_LASSO=rf.model_LASSO,svm_1=svm_Rad
ial_1,SVM_LASSO=svm_Radial_LASSO))
summary(CV)
```

```
##
## Call:
## summary.resamples(object = CV)
##
## Models: LR_1, LR_LASSO, RF_1, RF_LASSO, svm_1, SVM_LASSO
## Number of resamples: 10
##
## Accuracy
##             Min.  1st Qu.  Median   Mean 3rd Qu.   Max. NA's
## LR_1       0.7753  0.8017 0.8212 0.8238  0.8511 0.8652    0
## LR_LASSO   0.7416  0.7730 0.8045 0.8047  0.8315 0.8764    0
## RF_1       0.7640  0.8090 0.8248 0.8249  0.8404 0.8989    0
## RF_LASSO   0.7978  0.8022 0.8146 0.8227  0.8422 0.8539    0
## svm_1      0.7778  0.8095 0.8258 0.8272  0.8422 0.8764    0
## SVM_LASSO 0.7528  0.8090 0.8212 0.8227  0.8471 0.8764    0
##
## Kappa
##             Min.  1st Qu.  Median   Mean 3rd Qu.   Max. NA's
## LR_1       0.5241  0.5742 0.6189 0.6236  0.6752 0.7207    0
## LR_LASSO   0.4496  0.5187 0.5817 0.5834  0.6348 0.7426    0
## RF_1       0.4974  0.5861 0.6267 0.6235  0.6592 0.7796    0
## RF_LASSO   0.5351  0.5642 0.5853 0.6060  0.6527 0.6853    0
## svm_1      0.5122  0.5811 0.6147 0.6219  0.6545 0.7307    0
## SVM_LASSO 0.4582  0.5801 0.5986 0.6066  0.6596 0.7338    0
```

```
#Visualize
dotplot(CV)
```




```
#write.csv(pred_SVM_LASSO,"D:/LearningR/Titanic/perdiction_SVM_LASSO.csv")
```

d. 神经网络

```
set.seed(123)
#create dummy columns for caregorical variables
xfactors<-model.matrix(~Pclass+Sex+Survived+Title+Age+SibSp+Parch+Fare+Embarked,data=complete_data)
head(complete_data)
```

```
##   X Pclass    Sex Age SibSp Parch    Fare Cabin Embarked Title Survived
## 1 1      3   male  22     1     0  7.2500              S    Mr        0
## 2 2      1 female  38     1     0 71.2833   C85        C   Mrs        1
## 3 3      3 female  26     0     0  7.9250              S  Miss        1
## 4 4      1 female  35     1     0 53.1000  C123        S   Mrs        1
## 5 5      3   male  35     0     0  8.0500              S    Mr        0
## 6 6      3   male  30     0     0  8.4583              Q    Mr        0
```

```
complete_data<-data.frame(xfactors)

head(complete_data)
```

```
##   X.Intercept. Pclass Sexmale Survived1 TitleCol TitleCountess TitleDon
## 1            1      3       1         0        0             0        0
## 2            1      1       0         1        0             0        0
## 3            1      3       0         1        0             0        0
## 4            1      1       0         1        0             0        0
## 5            1      3       1         0        0             0        0
## 6            1      3       1         0        0             0        0
##   TitleDona TitleDr TitleJonkheer TitleLady TitleMajor TitleMaster
## 1         0       0             0         0          0           0
## 2         0       0             0         0          0           0
## 3         0       0             0         0          0           0
## 4         0       0             0         0          0           0
## 5         0       0             0         0          0           0
## 6         0       0             0         0          0           0
##   TitleMiss TitleMlle TitleMme TitleMr TitleMrs TitleMs TitleRev TitleSir
## 1         0         0        0       1        0       0        0        0
## 2         0         0        0       0        1       0        0        0
## 3         1         0        0       0        0       0        0        0
## 4         0         0        0       0        1       0        0        0
## 5         0         0        0       1        0       0        0        0
## 6         0         0        0       1        0       0        0        0
##   Age SibSp Parch    Fare EmbarkedC EmbarkedQ EmbarkedS
## 1  22     1     0  7.2500         0         0         1
## 2  38     1     0 71.2833         1         0         0
## 3  26     0     0  7.9250         0         0         1
## 4  35     1     0 53.1000         0         0         1
## 5  35     0     0  8.0500         0         0         1
## 6  30     0     0  8.4583         0         1         0
```

```
new_train<-complete_data[c(1:891),]
new_test<-complete_data[c(892:1309),]
str(new_train)
```

```
## 'data.frame':    891 obs. of  28 variables:
##  $ X.Intercept. : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Pclass       : num  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sexmale      : num  1 0 0 0 1 1 1 1 0 0 ...
##  $ Survived1    : num  0 1 1 1 0 0 0 0 1 1 ...
##  $ TitleCol     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleCountess: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleDon     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleDona    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleDr      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleJonkheer: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleLady    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleMajor   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleMaster  : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ TitleMiss    : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ TitleMlle    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleMme     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleMr      : num  1 0 0 0 1 1 1 0 0 0 ...
##  $ TitleMrs     : num  0 1 0 1 0 0 0 0 1 1 ...
##  $ TitleMs      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleRev     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TitleSir     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Age          : num  22 38 26 35 35 30 54 2 27 14 ...
##  $ SibSp        : num  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch        : num  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare         : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ EmbarkedC    : num  0 1 0 0 0 0 0 0 0 1 ...
##  $ EmbarkedQ    : num  0 0 0 0 0 1 0 0 0 0 ...
##  $ EmbarkedS    : num  1 0 1 1 1 0 1 1 1 0 ...
```

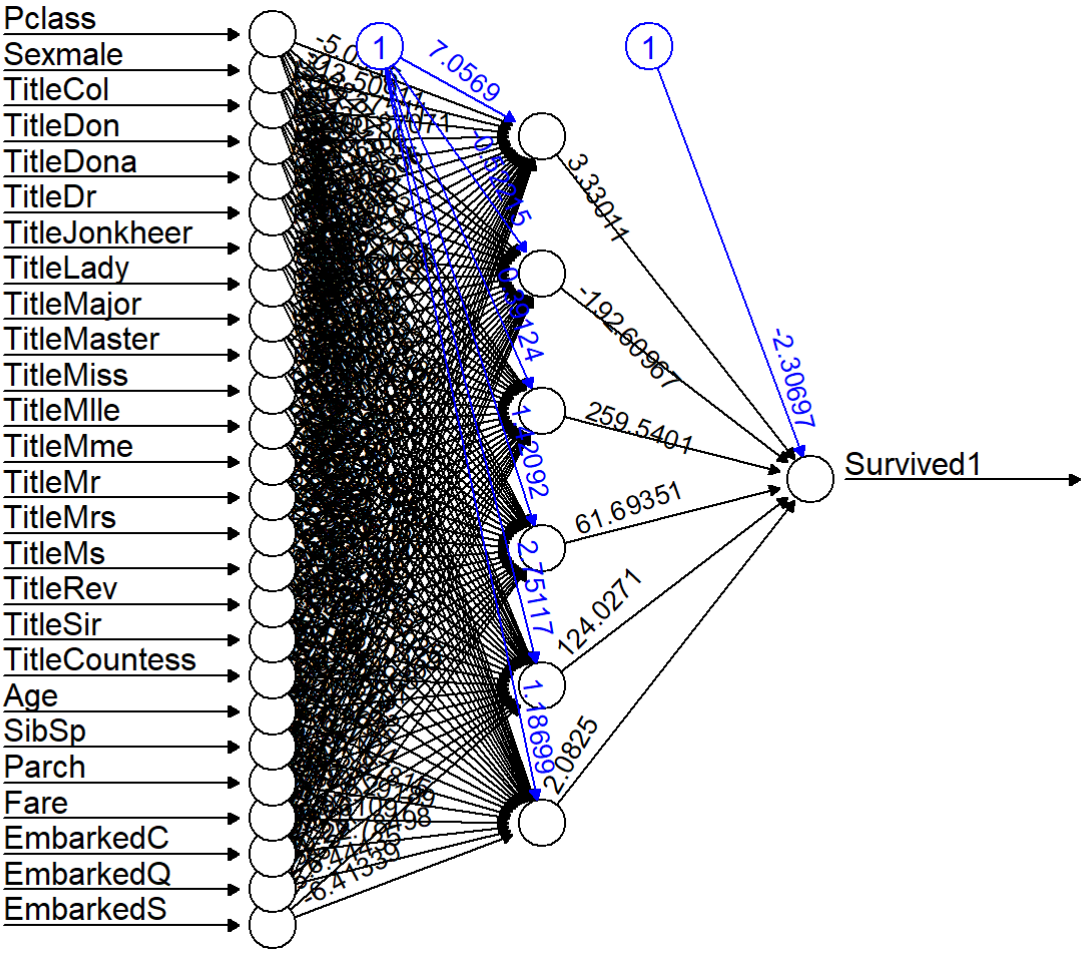
```
head(new_train)
```

```
##   X.Intercept. Pclass Sexmale Survived1 TitleCol TitleCountess TitleDon
## 1            1      3       1         0        0             0        0
## 2            1      1       0         1        0             0        0
## 3            1      3       0         1        0             0        0
## 4            1      1       0         1        0             0        0
## 5            1      3       1         0        0             0        0
## 6            1      3       1         0        0             0        0
##   TitleDona TitleDr TitleJonkheer TitleLady TitleMajor TitleMaster
## 1         0       0             0         0          0           0
## 2         0       0             0         0          0           0
## 3         0       0             0         0          0           0
## 4         0       0             0         0          0           0
## 5         0       0             0         0          0           0
## 6         0       0             0         0          0           0
##   TitleMiss TitleMlle TitleMme TitleMr TitleMrs TitleMs TitleRev TitleSir
## 1         0         0        0       1        0       0        0        0
## 2         0         0        0       0        1       0        0        0
## 3         1         0        0       0        0       0        0        0
## 4         0         0        0       0        1       0        0        0
## 5         0         0        0       1        0       0        0        0
## 6         0         0        0       1        0       0        0        0
##   Age SibSp Parch    Fare EmbarkedC EmbarkedQ EmbarkedS
## 1  22     1     0  7.2500         0         0         1
## 2  38     1     0 71.2833         1         0         0
## 3  26     0     0  7.9250         0         0         1
## 4  35     1     0 53.1000         0         0         1
## 5  35     0     0  8.0500         0         0         1
## 6  30     0     0  8.4583         0         1         0
```

```
set.seed(123)
NN_1 <- neuralnet(Survived1 ~Pclass+Sexmale     +TitleCol   +TitleDon+  TitleDona   +TitleDr    +Title
Jonkheer+ TitleLady   +TitleMajor +TitleMaster+   TitleMiss   +TitleMlle  +TitleMme   +TitleMr    +Tit
leMrs+  TitleMs+    TitleRev    +TitleSir+  TitleCountess   +Age+   SibSp   +Parch+ Fare
+EmbarkedC+ EmbarkedQ   +EmbarkedS, new_train, hidden =6, lifesign = "minimal",     linear.output = FA
LSE, threshold = 0.1)
```

```
## hidden: 6    thresh: 0.1    rep: 1/1    steps:    6588   error: 42.29669 time: 9.34 secs
```

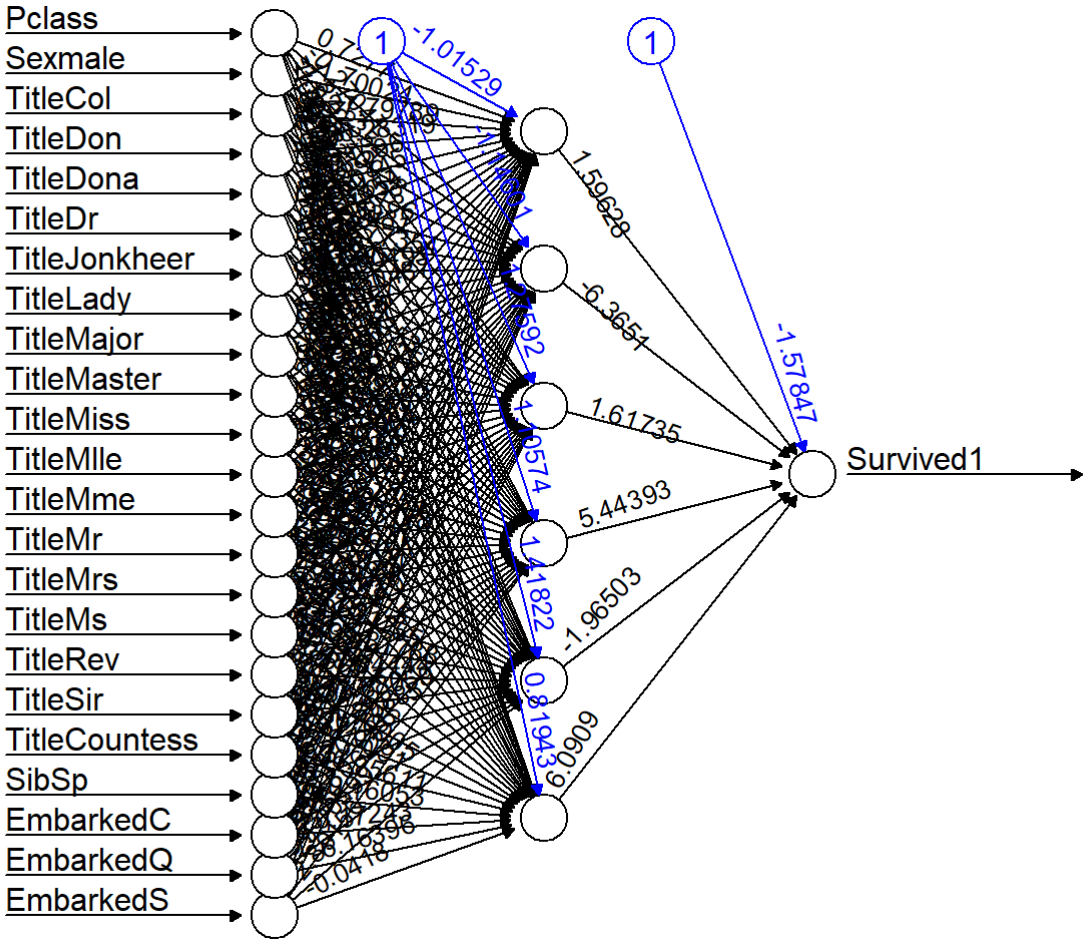
```
#plot
plot(NN_1, rep = "best")
```




```
#predict
output <- compute(NN_1, new_test[,-c(1,4)],rep=1)
predict<- output$net.result
pred_NN<-round(predict,digit=0)
#write.csv(pred_NN,"D:/LearningR/Titanic/perdiction_NN.csv")

#使用lasso回归结果
set.seed(123)
NN_LASSO <- neuralnet(Survived1 ~Pclass+Sexmale     +TitleCol   +TitleDon+  TitleDona   +TitleDr    +T
itleJonkheer+ TitleLady   +TitleMajor +TitleMaster+   TitleMiss   +TitleMlle  +TitleMme   +TitleMr
+TitleMrs+  TitleMs+    TitleRev    +TitleSir+  TitleCountess   +   SibSp       +EmbarkedC+ EmbarkedQ
  +EmbarkedS, new_train, hidden =6, lifesign = "minimal",     linear.output = FALSE, threshold = 0.1)
```

```
## hidden: 6    thresh: 0.1    rep: 1/1    steps:     383    error: 50.50515 time: 0.6 secs
```

```
#plot
plot(NN_LASSO, rep = "best")
```




```
#predict
output <- compute(NN_LASSO, new_test[,-c(1,4,22,24,25)],rep=1)
predict<- output$net.result
pred_NN_LASSO<-round(predict,digit=0)
#write.csv(pred_NN_LASSO,"D:/LearningR/Titanic/perdiction_NN_LASSO.csv")
```

结果分析： 采用的预测模型有：逻辑回归，随机森林，支持向量机，神经网络 lasso特征选择的结果作为第二个基本模型（2），与将全部变量放入模型（1）的结果进行比较

交叉验证结果， 支持向量机（1）给出的平均准确性最高，随机森林（2）最为稳健

预测结果准确度： SVM_1:0.799 SVM_LASSO: 0.7942 Randomforest_1:0.7846 Random_Forest_LASSO: 0.7799 Logistic_LASSO:0.77512 神经网络_1：0.6266 神经网络__LASSO:0.6028

局限性： 1. 在使用神经网络和支持向量机的时候参数选择比较粗糙，由于时间限制未对Age Fare标准化,hiddenlayer的个数并未逐一尝试，由于缺乏专业知识，支持向量机kernel的选择是使用一般默认的Radial basis function 等诸多问题 ，导致模型预测准确性很低 2. 变量选择只用了lasso回归，可能有其他更合适的方法进行变量选择