incomepredict

yayun

2023-01-05

导入包

```
library(caret)

## 载入需要的程辑包：ggplot2

## 载入需要的程辑包：lattice

library(dplyr)

##
## 载入程辑包：'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(knitr)
library(ROCR)
library(tinytex)
```

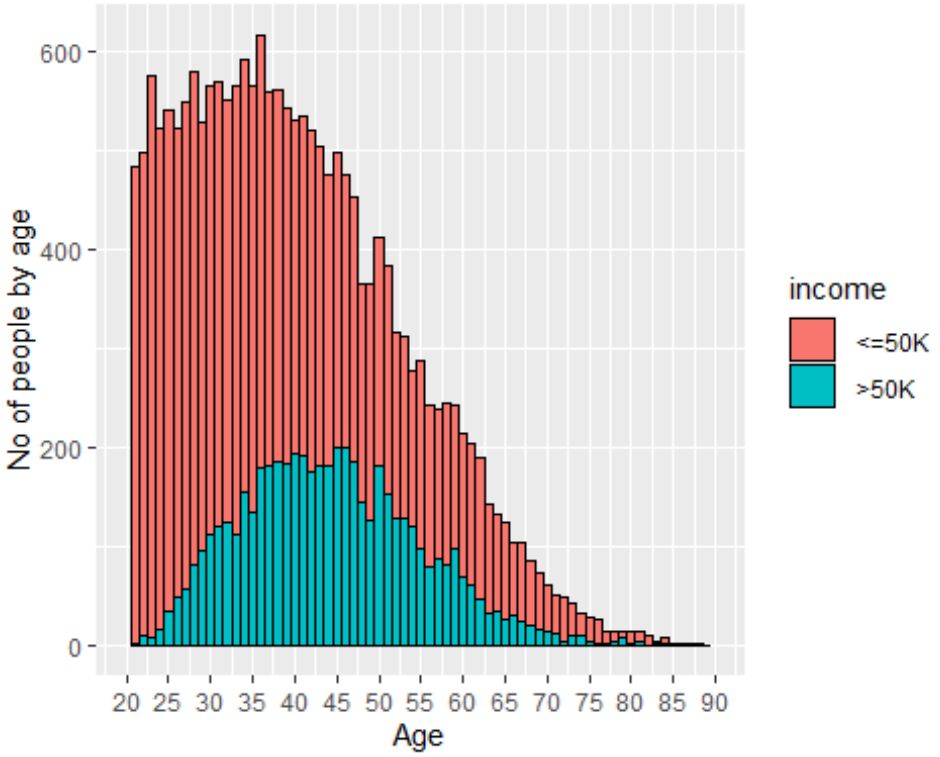
数据准备

```
#载入数据
data <- read.table("D://R-data/Income.data", header = F, sep = ",", na.strings = " ?")
#给列命名
names(data) <- c("age","workclass","fnlwgt","education","educational.num","marital.status","Occupation","Relationship","race","Sex","capital.gain","capital.loss","Hours.per.week","native.country","income")
#查看数据类型
str(data)

## 'data.frame':    21531 obs. of  15 variables:
##  $ age            : int  27 38 33 43 36 38 24 22 30 55 ...
##  $ workclass      : chr  " Self-emp-not-inc" " Private" " Private" " Private" ...
##  $ fnlwgt         : int  41099 472604 348618 135606 248445 112093 197552 303822 288566 487411 ...
```

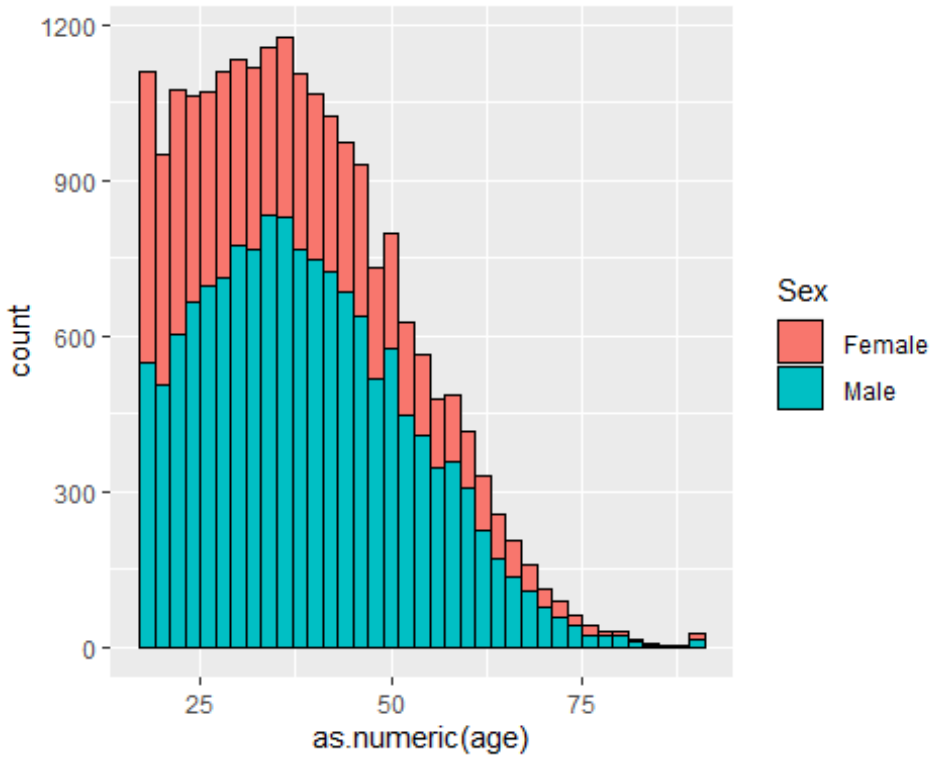
```
##  $ education      : chr  " HS-grad" " Bachelors" " 5th-6th" " Master
s" ...
##  $ educational.num: int  9 13 3 14 9 7 9 6 13 14 ...
##  $ marital.status : chr  " Never-married" " Married-civ-spouse" " Ma
rried-spouse-absent" " Divorced" ...
##  $ Occupation     : chr  " Craft-repair" " Other-service" " Transpor
t-moving" " Exec-managerial" ...
##  $ Relationship   : chr  " Not-in-family" " Husband" " Unmarried" "
Not-in-family" ...
##  $ race           : chr  " White" " White" " Other" " White" ...
##  $ Sex            : chr  " Male" " Male" " Male" " Female" ...
##  $ capital.gain   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hours.per.week : int  30 35 20 50 60 40 40 40 40 40 ...
##  $ native.country : chr  " United-States" " Mexico" " El-Salvador" "
 United-States" ...
##  $ income         : chr  " <=50K" " <=50K" " <=50K" " >50K" ...
```

探索数据

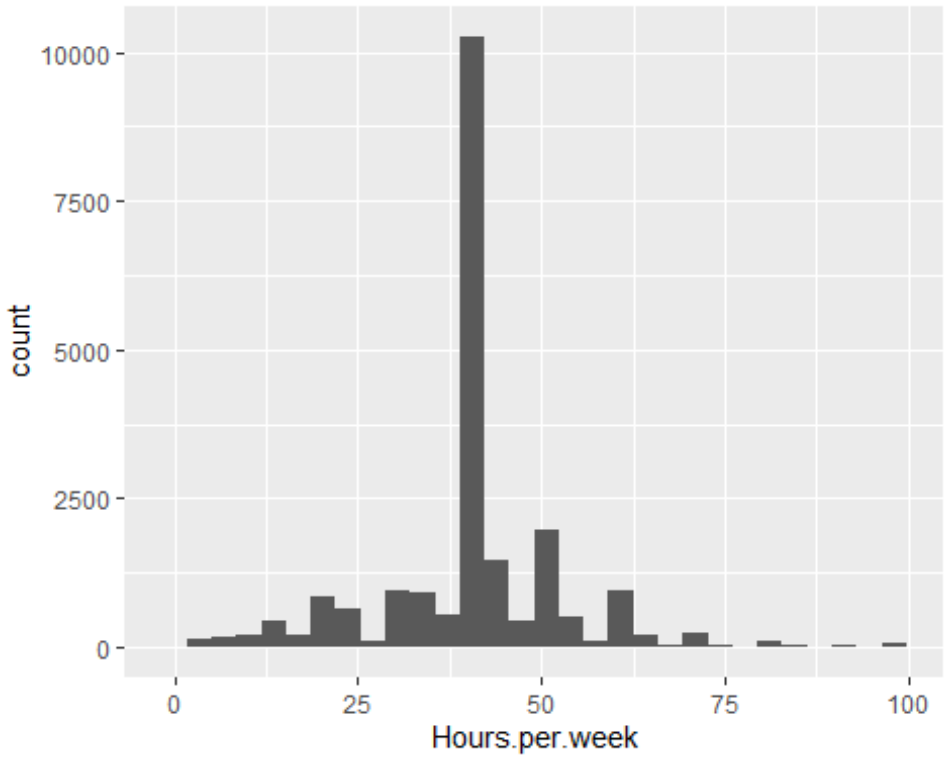
```
library(ggplot2)
#查看年龄和收入的柱状图
ggplot(data, aes(age)) + geom_histogram(aes(fill = income), color = "bl
ack",binwidth = 1)+
  scale_x_continuous(limits=c(20,90), breaks=seq(20,90,5)) +
  xlab("Age") +
  ylab("No of people by age")

## Warning: Removed 1110 rows containing non-finite values (`stat_bin()
`).

## Warning: Removed 4 rows containing missing values (`geom_bar()`).
```




```
#结论：收入大于50K 的人群年龄集中在30-50 岁之间

#查看年龄和性别的柱状图
ggplot(data) + aes(x=as.numeric(age),group=Sex, fill=Sex) +geom_histogr
am(binwidth=2,color='black')
```

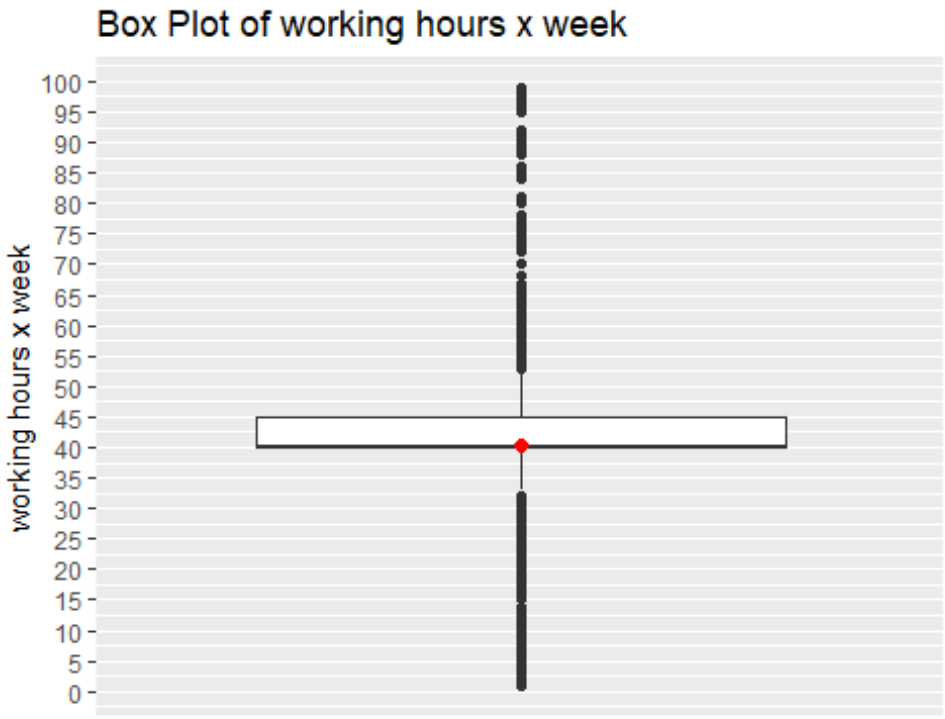



```
#结论：数据中男性占比更大，年龄集中在 20-50 岁

#查看每周工作小时数
ggplot(data, aes(Hours.per.week)) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

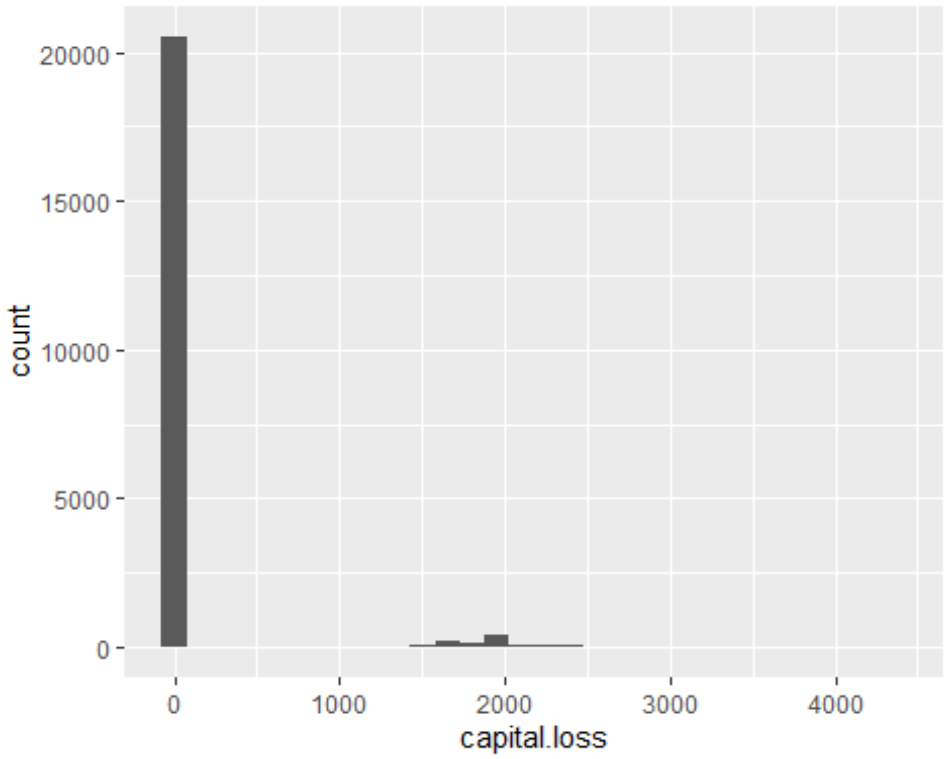



```
#结论：最高频的数据是每周工作40h（平均每天工作八小时，很良心）

#更直观地方法：箱型图输出
ggplot(aes(x = factor(0), y = Hours.per.week),
       data = data) +
  geom_boxplot() +
  stat_summary(fun.y = mean,
               geom = 'point',
               shape = 19,
               color = "red",
               cex = 2) +
  scale_x_discrete(breaks = NULL) +
  scale_y_continuous(breaks = seq(0, 100, 5)) +
  xlab(label = "") +
  ylab(label = "working hours x week") +
  ggtitle("Box Plot of working hours x week")
```

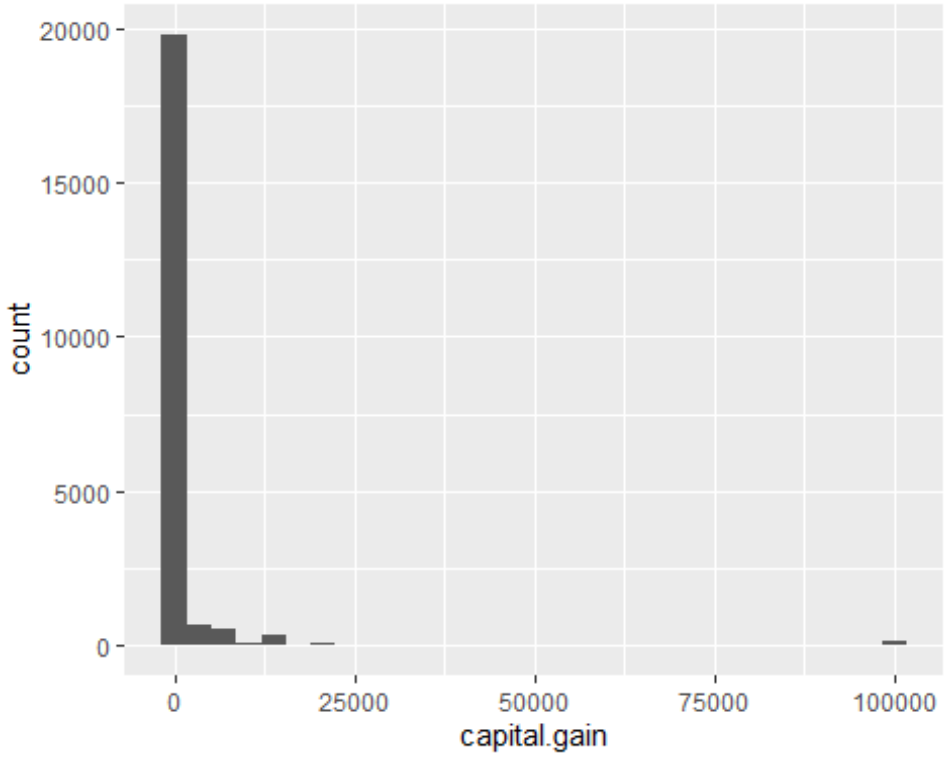
```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as o
f ggplot2 3.3.0.
## ℹ Please use the `fun` argument instead.
```




```
#结论：平均工作小时数为40 标记为箱线图上的红点，并且至少50%参与调查的人，每周工作40 至45 小时

#查看资本损失
ggplot(data, aes(capital.loss)) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#结论：资本损失可能对分类没有用，因为它非常倾斜并且主要集中在零值

#查看资本增益
ggplot(data, aes(capital.gain)) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#结论：资本增益可能对分类没有用，因为它非常倾斜并且主要集中在零值
```

数据处理

```
sapply(data, function(x) sum(is.na(x)))
```

```
##             age       workclass          fnlwgt       education educ
ational.num
##               0            1192               0               0
          0
##  marital.status      Occupation    Relationship            race
        Sex
##               0            1197               0               0
          0
##    capital.gain    capital.loss  Hours.per.week  native.country
     income
##               0               0               0             385
          0
```

```
#缺失值处理，填补有缺失值的部分
data$workclass[is.na(data$workclass)]   <- " Private"
data$Occupation[is.na(data$Occupation)] <- " Prof-Specialty"
data$native.country[is.na(data$native.country)]       <- " United-State
s"
sapply(data, function(x) sum(is.na(x)))
```

```
##             age       workclass          fnlwgt       education educ
ational.num
```

```
##              0               0               0               0
##              0
##  marital.status      Occupation    Relationship            race
##             Sex
##              0               0               0               0
##              0
##    capital.gain    capital.loss  Hours.per.week  native.country
##          income
##              0               0               0               0
##              0
```

```
#异常值处理

Outlier1 <- boxplot (data$age, plot = FALSE)$out
data <- data[-which(data$age %in% Outlier1 ),]

Outlier2 <- boxplot (data$fnlwgt, plot = FALSE)$out
data <- data[-which(data$fnlwgt %in% Outlier2 ),]

Outlier3 <- boxplot (data$educational.num, plot = FALSE)$out
data <- data[-which(data$educational.num %in% Outlier3 ),]

Outlier6 <- boxplot (data$Hours.per.week, plot = FALSE)$out
data <- data[-which(data$Hours.per.week %in% Outlier6 ),]

#将收入是否大于50K 转换为0/1 表示  >50K
data$income<-ifelse(data$income==' >50K',1,0)

#将 chr 型改为factor 型
data <- data %>%
  mutate_if(is.character,as.factor)
str(data)
```

```
## 'data.frame':    14533 obs. of  15 variables:
##  $ age            : int  43 38 24 22 30 39 50 34 46 45 ...
##  $ workclass      : Factor w/ 8 levels " Federal-gov",..: 4 4 2 4 4
7 4 4 4 4 ...
##  $ fnlwgt         : int  135606 112093 197552 303822 288566 239409 3
37606 32528 234690 190482 ...
##  $ education      : Factor w/ 12 levels " 10th"," 11th",..: 10 2 9 1
 7 9 9 6 9 9 ...
##  $ educational.num: int  14 7 9 6 13 9 9 11 9 9 ...
##  $ marital.status : Factor w/ 7 levels " Divorced"," Married-AF-spou
se",..: 1 3 5 5 5 3 3 4 3 1 ...
##  $ Occupation     : Factor w/ 15 levels " Adm-clerical",..: 4 15 14
3 8 12 7 1 13 7 ...
##  $ Relationship   : Factor w/ 6 levels " Husband"," Not-in-family
",..: 2 1 2 4 2 1 1 5 1 5 ...
##  $ race           : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5
 5 5 5 5 5 5 5 ...
```

```
##  $ Sex            : Factor w/ 2 levels " Female"," Male": 1 2 1 2 2
2 2 1 2 2 ...
##  $ capital.gain   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss   : int  0 0 0 0 0 0 1485 974 0 0 ...
##  $ Hours.per.week : int  50 40 40 40 40 50 40 40 40 40 ...
##  $ native.country : Factor w/ 41 levels " Cambodia"," Canada",..: 39
 39 39 39 39 39 39 39 39 39 ...
##  $ income         : num  1 0 0 0 0 0 0 0 0 0 ...
```

建模

```
#划分测试集训练集
library(caTools)
split <- sample.split(data$income, SplitRatio = 0.7)
#因为在几次尝试随机划分时，发现如果Holand-Netherlands 没被划分到训练集，预测
时层数不匹配就会报错，所以训练集中强制加入了该数据
train <- subset(data, split == TRUE | native.country==' Holand-Netherla
nds')
test <- subset(data, split == FALSE)
#10 折交叉验证
train.control <- trainControl(method="cv",number=10)
log.model <- glm(income ~ ., family = binomial(), train)
```

查看系数关系

```
# 查看模型信息
print(log.model)

##
## Call:  glm(formula = income ~ ., family = binomial(), data = train)
##
## Coefficients:
##                                 (Intercept)
##                                   -8.848e+00
##                                          age
##                                    3.649e-02
##                           workclass Local-gov
##                                   -7.910e-01
##                        workclass Never-worked
##                                   -1.461e+01
##                             workclass Private
##                                   -6.038e-01
##                        workclass Self-emp-inc
##                                   -4.180e-01
##                     workclass Self-emp-not-inc
##                                   -1.346e+00
##                           workclass State-gov
##                                   -9.123e-01
##                         workclass Without-pay
##                                   -1.541e+01
##                                       fnlwgt
```

```
##                                 1.286e-06
##                            education 11th
##                                 1.767e-01
##                            education 12th
##                                -2.916e-02
##                             education 9th
##                                -1.077e-01
##                       education Assoc-acdm
##                                 1.539e+00
##                        education Assoc-voc
##                                 1.351e+00
##                        education Bachelors
##                                 2.091e+00
##                        education Doctorate
##                                 3.164e+00
##                          education HS-grad
##                                 7.617e-01
##                          education Masters
##                                 2.408e+00
##                       education Prof-school
##                                 2.769e+00
##                     education Some-college
##                                 1.235e+00
##                             educational.num
##                                          NA
##          marital.status Married-AF-spouse
##                                 1.788e+01
##         marital.status Married-civ-spouse
##                                 2.523e+00
##      marital.status Married-spouse-absent
##                                -3.416e-01
##               marital.status Never-married
##                                -4.077e-01
##                   marital.status Separated
##                                -2.494e-01
##                     marital.status Widowed
##                                 3.196e-01
##                    Occupation Armed-Forces
##                                -1.192e+00
##                    Occupation Craft-repair
##                                 5.153e-02
##                  Occupation Exec-managerial
##                                 6.485e-01
##                  Occupation Farming-fishing
##                                -1.139e+00
##               Occupation Handlers-cleaners
##                                -7.144e-01
##               Occupation Machine-op-inspct
##                                -2.811e-01
##                      Occupation Other-service
```

```
##                                -1.034e+00
##                Occupation Priv-house-serv
##                                -3.773e+00
##                 Occupation Prof-specialty
##                                 4.758e-01
##                 Occupation Prof-Specialty
##                                -1.137e+00
##                Occupation Protective-serv
##                                 6.387e-01
##                          Occupation Sales
##                                 2.273e-01
##                   Occupation Tech-support
##                                 4.481e-01
##               Occupation Transport-moving
##                                -2.290e-01
##                Relationship Not-in-family
##                                 7.444e-01
##               Relationship Other-relative
##                                -2.408e-01
##                    Relationship Own-child
##                                -2.036e-02
##                    Relationship Unmarried
##                                 6.458e-01
##                         Relationship Wife
##                                 1.336e+00
##                   race Asian-Pac-Islander
##                                 1.087e+00
##                                race Black
##                                 7.172e-01
##                                race Other
##                                 1.008e+00
##                                race White
##                                 9.355e-01
##                                  Sex Male
##                                 9.743e-01
##                             capital.gain
##                                 3.575e-04
##                             capital.loss
##                                 6.449e-04
##                           Hours.per.week
##                                 4.942e-02
##                     native.country Canada
##                                -1.932e+00
##                      native.country China
##                                -9.542e-01
##                    native.country Columbia
##                                -1.512e+01
##                       native.country Cuba
##                                -6.860e-01
##         native.country Dominican-Republic
```

```
##                                  -1.445e+01
##                     native.country Ecuador
##                                  -1.558e+00
##                native.country El-Salvador
##                                  -9.779e-01
##                     native.country England
##                                  -1.259e-01
##                      native.country France
##                                   6.112e-01
##                     native.country Germany
##                                  -4.084e-01
##                      native.country Greece
##                                  -4.878e+00
##                   native.country Guatemala
##                                  -1.474e+01
##                        native.country Haiti
##                                  -1.067e+00
##         native.country Holand-Netherlands
##                                  -1.296e+01
##                    native.country Honduras
##                                  -1.425e+01
##                       native.country Hong
##                                  -2.101e+00
##                     native.country Hungary
##                                   2.895e+00
##                       native.country India
##                                  -1.237e+00
##                        native.country Iran
##                                  -1.302e+00
##                     native.country Ireland
##                                   2.849e-01
##                       native.country Italy
##                                   7.230e-02
##                     native.country Jamaica
##                                  -2.229e+00
##                       native.country Japan
##                                  -1.226e-01
##                        native.country Laos
##                                  -1.685e+00
##                      native.country Mexico
##                                  -1.605e+00
##                   native.country Nicaragua
##                                  -9.952e-01
## native.country Outlying-US(Guam-USVI-etc)
##                                  -1.565e+01
##                        native.country Peru
##                                  -1.474e+01
##                native.country Philippines
##                                  -3.599e-01
##                      native.country Poland
```

```
##                                   -8.262e-01
##                     native.country Portugal
##                                   -1.741e+00
##                  native.country Puerto-Rico
##                                   -1.368e+00
##                     native.country Scotland
##                                   -4.960e-01
##                        native.country South
##                                   -1.270e+00
##                       native.country Taiwan
##                                   -5.294e-01
##                     native.country Thailand
##                                    4.227e-01
##              native.country Trinadad&Tobago
##                                    2.143e+00
##                native.country United-States
##                                   -6.102e-01
##                       native.country Vietnam
##                                   -1.292e+00
##                    native.country Yugoslavia
##                                   -5.705e-01
##
## Degrees of Freedom: 10173 Total (i.e. Null);  10080 Residual
## Null Deviance:       11730
## Residual Deviance: 6718  AIC: 6906
```

```
#用kable的方式更直观
kable(summary(log.model)$coefficients, longtable=TRUE, digits = 6)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.848111	1.307123	-6.769152	0.000000
age	0.036489	0.003069	11.888325	0.000000
workclass Local-gov	-0.790997	0.184614	-4.284590	0.000018
workclass Never-worked	-14.607497	2399.544739	-0.006088	0.995143
workclass Private	-0.603828	0.150563	-4.010459	0.000061
workclass Self-emp-inc	-0.417955	0.215536	-1.939146	0.052484
workclass Self-emp-not-inc	-1.346200	0.191835	-7.017482	0.000000
workclass State-gov	-0.912273	0.207510	-4.396277	0.000011
workclass Without-pay	-15.408029	1120.263146	-0.013754	0.989026
fnlwgt	0.000001	0.000000	3.581832	0.000341
education 11th	0.176683	0.359310	0.491730	0.622910
education 12th	-0.029165	0.465033	-0.062715	0.949993
education 9th	-0.107680	0.446688	-0.241063	0.809506

	Estimate	Std. Error	z value	Pr(>|z|)
education Assoc-acdm	1.539017	0.301767	5.100018	0.000000
education Assoc-voc	1.350575	0.290580	4.647862	0.000003
education Bachelors	2.090694	0.270947	7.716250	0.000000
education Doctorate	3.163838	0.390869	8.094360	0.000000
education HS-grad	0.761693	0.262607	2.900507	0.003726
education Masters	2.407795	0.291222	8.267895	0.000000
education Prof-school	2.768749	0.351849	7.869142	0.000000
education Some-college	1.235485	0.266916	4.628745	0.000004
marital.status Married-AF-spouse	17.883123	1292.237852	0.013839	0.988959
marital.status Married-civ-spouse	2.523278	0.439119	5.746230	0.000000
marital.status Married-spouse-absent	-0.341554	0.422082	-0.809212	0.418393
marital.status Never-married	-0.407749	0.154214	-2.644042	0.008192
marital.status Separated	-0.249425	0.292598	-0.852450	0.393964
marital.status Widowed	0.319632	0.271292	1.178187	0.238722
Occupation Armed-Forces	-1.191861	1.774949	-0.671490	0.501908
Occupation Craft-repair	0.051530	0.133585	0.385750	0.699682
Occupation Exec-managerial	0.648529	0.130365	4.974706	0.000001
Occupation Farming-fishing	-1.139202	0.307717	-3.702104	0.000214
Occupation Handlers-cleaners	-0.714387	0.240554	-2.969765	0.002980
Occupation Machine-op-inspct	-0.281057	0.167180	-1.681162	0.092731
Occupation Other-service	-1.034240	0.218472	-4.733972	0.000002
Occupation Priv-house-serv	-3.773464	3.077743	-1.226049	0.220180
Occupation Prof-specialty	0.475791	0.137756	3.453860	0.000553
Occupation Prof-Specialty	-1.137169	0.235654	-4.825581	0.000001
Occupation Protective-serv	0.638730	0.218471	2.923632	0.003460
Occupation Sales	0.227328	0.142486	1.595445	0.110613
Occupation Tech-support	0.448059	0.181968	2.462301	0.013805
Occupation Transport-moving	-0.229036	0.175266	-1.306790	0.191284
Relationship Not-in-family	0.744425	0.434088	1.714915	0.086361
Relationship Other-relative	-0.240809	0.448815	-0.536543	0.591583

	Estimate	Std. Error	z value	Pr(>|z|)
Relationship Own-child	-0.020363	0.415878	-0.048965	0.960947
Relationship Unmarried	0.645759	0.463722	1.392558	0.163754
Relationship Wife	1.336024	0.181987	7.341304	0.000000
race Asian-Pac-Islander	1.087189	0.484664	2.243180	0.024885
race Black	0.717186	0.435879	1.645377	0.099892
race Other	1.008239	0.588905	1.712058	0.086886
race White	0.935542	0.418505	2.235439	0.025389
Sex Male	0.974329	0.139373	6.990785	0.000000
capital.gain	0.000358	0.000019	18.993339	0.000000
capital.loss	0.000645	0.000065	9.929535	0.000000
Hours.per.week	0.049420	0.007715	6.405833	0.000000
native.country Canada	-1.931611	1.267612	-1.523819	0.127554
native.country China	-0.954198	1.162439	-0.820859	0.411727
native.country Columbia	-15.119654	380.536664	-0.039732	0.968306
native.country Cuba	-0.686037	1.155112	-0.593914	0.552570
native.country Dominican-Republic	-14.453483	429.410209	-0.033659	0.973149
native.country Ecuador	-1.558004	1.580897	-0.985519	0.324369
native.country El-Salvador	-0.977907	1.259949	-0.776148	0.437662
native.country England	-0.125931	1.165928	-0.108009	0.913988
native.country France	0.611216	1.335669	0.457610	0.647233
native.country Germany	-0.408449	1.097396	-0.372198	0.709745
native.country Greece	-4.878037	3.210122	-1.519580	0.128617
native.country Guatemala	-14.741098	544.196027	-0.027088	0.978390
native.country Haiti	-1.067291	1.610582	-0.662674	0.507539
native.country Holand-Netherlands	-12.957967	2399.544970	-0.005400	0.995691
native.country Honduras	-14.246057	1330.301165	-0.010709	0.991456
native.country Hong	-2.101220	1.501581	-1.399339	0.161711
native.country Hungary	2.895326	5.491254	0.527261	0.598012
native.country India	-1.237355	1.077284	-1.148588	0.250726
native.country Iran	-1.302394	1.233759	-1.055630	0.291137
native.country Ireland	0.284903	1.448614	0.196673	0.844084

	Estimate	Std. Error	z value	Pr(>|z|)
native.country Italy	0.072295	1.253874	0.057658	0.954021
native.country Jamaica	-2.228759	1.509460	-1.476528	0.139802
native.country Japan	-0.122592	1.201320	-0.102048	0.918719
native.country Laos	-1.685109	1.603020	-1.051209	0.293163
native.country Mexico	-1.604709	1.124767	-1.426704	0.153665
native.country Nicaragua	-0.995247	1.526242	-0.652090	0.514343
native.country Outlying-US(Guam-USVI-etc)	-15.652608	982.057739	-0.015939	0.987283
native.country Peru	-14.742509	851.862271	-0.017306	0.986192
native.country Philippines	-0.359889	1.045216	-0.344320	0.730605
native.country Poland	-0.826177	1.326689	-0.622736	0.533458
native.country Portugal	-1.740581	1.814202	-0.959420	0.337347
native.country Puerto-Rico	-1.368028	1.185757	-1.153717	0.248616
native.country Scotland	-0.496042	1.430469	-0.346769	0.728765
native.country South	-1.270243	1.238107	-1.025956	0.304912
native.country Taiwan	-0.529446	1.320620	-0.400907	0.688489
native.country Thailand	0.422710	3.418663	0.123648	0.901594
native.country Trinadad&Tobago	2.142529	1.626660	1.317134	0.187794
native.country United-States	-0.610230	1.017984	-0.599450	0.548873
native.country Vietnam	-1.291631	1.310314	-0.985741	0.324260
native.country Yugoslavia	-0.570539	1.528787	-0.373197	0.709002

列分别表示：估值,标准误差,T 值,P 值。因为 P 值估计系数不显著的可能性，有较大 P 值的变量是可以从模型中移除的候选变量。可以看到 native.country 的 p 值大多数符合以上所说，之后可以考虑去掉这个特征。

模型评估

```
#混淆矩阵
cm<-with(train,table(test$income, prediction >= 0.5))
cm

##
##     FALSE TRUE
##   0  2939  272
##   1   451  698
```

```r
#精度
trainacc<- (cm[1,1]+cm[2,2])/(cm[1,1]+cm[2,2]+cm[1,2]+cm[2,1])
trainacc
```

```
## [1] 0.8341743
```

```r
#准确率
trainprecision<- cm[2,2]/(cm[2,2]+cm[1,2])
trainprecision
```

```
## [1] 0.7195876
```

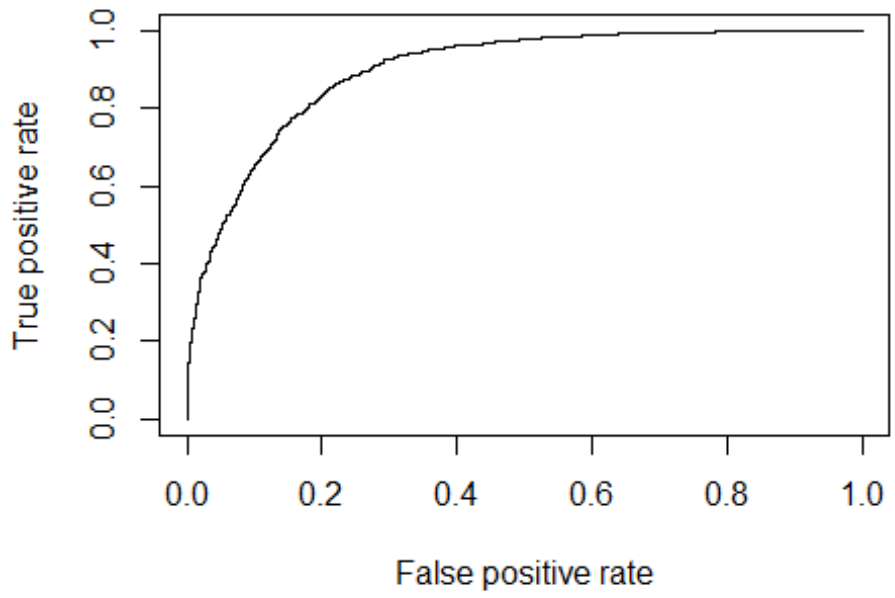
```r
#召回率
trainrecall<-cm[2,2]/(cm[2,2]+cm[2,1])
trainrecall
```

```
## [1] 0.6074848
```

```r
#F-score
trainfs<-2*(trainacc*trainrecall)/(trainacc+trainrecall)
trainfs
```

```
## [1] 0.703007
```

```r
#ROC 曲线
library(ROCR)
eval <- prediction(prediction,test$income)
plot(performance(eval,"tpr","fpr"))
```



```
#AUC 值
print(attributes(performance(eval,'auc'))$y.values[[1]])

## [1] 0.8978154
```

改进模型

```
#仔细观察data的14个属性，发现每一个属性下面有很多杂乱的值，可以进行一下合并
#合并工作阶层
data$workclass <- as.character(data$workclass)
data$workclass[data$workclass == " Without-pay" |
                 data$workclass == " Never-worked"] <- " Unemployed"

data$workclass[data$workclass == " State-gov" |
                 data$workclass == " Local-gov"] <- " SL-gov"

data$workclass[data$workclass == " Self-emp-inc" |
                 data$workclass == " Self-emp-not-inc"] <- " Self-emplo
yed"

table(data$workclass)

##
##    Federal-gov        Private  Self-employed         SL-gov     Unem
ployed
##            534          11026           1302           1667
     4

#合并婚姻状况
data$marital.status <- as.character(data$marital.status)

data$marital.status[data$marital.status == " Married-AF-spouse" |
                      data$marital.status == " Married-civ-spouse" |
                      data$marital.status == " Married-spouse-absent"]
<- " Married"

data$marital.status[data$marital.status == " Divorced" |
                      data$marital.status == " Separated" |
                      data$marital.status == " Widowed"] <- " Not-Marri
ed"
table(data$marital.status)

##
##        Married  Never-married    Not-Married
##           7200           4354           2979

#同理，合并受教育程度和职业两个属性
data$education <- as.character(data$education)
data$education[data$education == " 12th" |
                      data$education == " Preschool" |
                      data$education == " 1st_4th" |
```

```
                      data$education == " 10th" |
                      data$education == " 5th_6th"|
                      data$education == " 7th-8th" |
                      data$education == " 9th" |
                      data$educations == " 11th"] <- " Schooling"
data$education[data$education == " Bachelors" |
                 data$education == " HS_grad" |
                 data$education == " Some_college" ] <- " Graduate"
data$education[data$education == " Assoc_acdm" |
                 data$education == " Assoc_voc" |
                 data$education == " Prof_school" ] <- " Masters"

data$Occupation <- as.character(data$Occupation)
data$Occupation[data$Occupation == " Adm_clerical" |
                  data$Occupation == " Sales" |
                  data$Occupation == " Transport_moving" |
                  data$mOccupation == " Handlers_cleaners"] <- " Admin/
Clerical"
data$Occupation[data$Occupation == " Farming_fishing" |
                   data$Occupation == " Protective_serv" |
                   data$Occupation == " Armed_Forces" |
                   data$Occupation == " Other_service" |
                   data$mOccupation == " Priv_house_serv"] <- " others"
data$Occupation[data$Occupation == " Machine_op_inspct" |
                   data$Occupation == " Tech_support" ] <- " Technical"
data$Occupation[data$Occupation == " Exec_managerial" |
                 data$Occupation == " Prof_specialty" ] <- " Manageria
l"

#将 character 转回 factor
data <- data %>%
  mutate_if(is.character,as.factor)

#选择属性建模
log2.model <- glm(income~age+workclass+fnlwgt+education+educational.num
+marital.status+Occupation+Relationship+race+Sex+Hours.per.week+capita
l.gain+capital.loss, family = binomial(), train)

## Warning: glm.fit:拟合機率算出来是数值零或一

prediction <- predict(log2.model, test, type = "response")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if
(type == :
## 用秩缺乏拟合来进行预测的结果很可能不可靠

#混淆矩阵
cm<-with(train,table(test$income, prediction >= 0.5))
cm
```

```
##
##     FALSE TRUE
##   0  2940  271
##   1   460  689
```

```
#精度
trainacc<- (cm[1,1]+cm[2,2])/(cm[1,1]+cm[2,2]+cm[1,2]+cm[2,1])
trainacc
```

```
## [1] 0.8323394
```

```
#准确率
trainprecision<- cm[2,2]/(cm[2,2]+cm[1,2])
trainprecision
```

```
## [1] 0.7177083
```

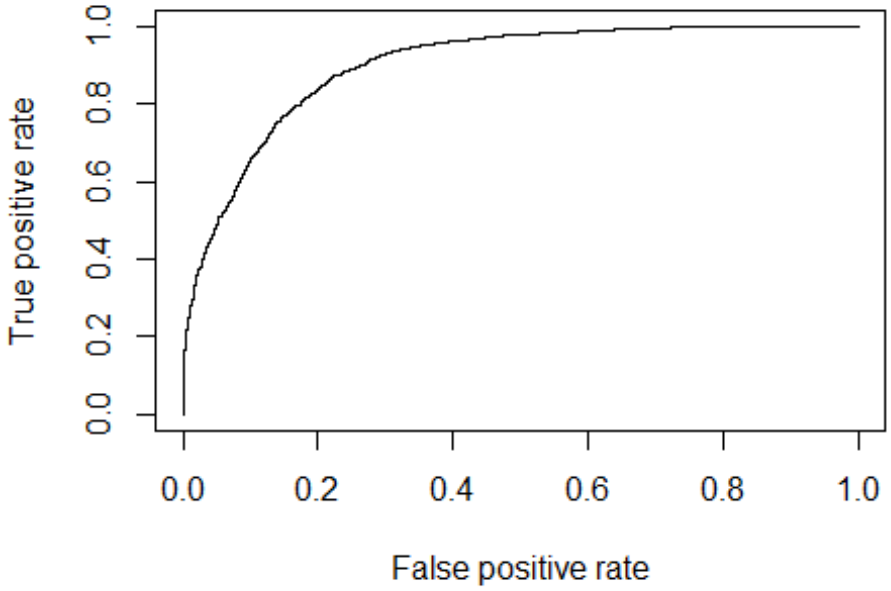
```
#召回率
trainrecall<-cm[2,2]/(cm[2,2]+cm[2,1])
trainrecall
```

```
## [1] 0.5996519
```

```
#F-score
trainfs<-2*(trainacc*trainrecall)/(trainacc+trainrecall)
trainfs
```

```
## [1] 0.6970907
```

```
#ROC 曲线
library(ROCR)
eval <- prediction(prediction,test$income)
plot(performance(eval,"tpr","fpr"))
```




```
#AUC 值
print(attributes(performance(eval,'auc'))$y.values[[1]])

## [1] 0.9099195
```

根据模型评估，发现数据提升并不明显，推测是因为数据量本身不丰富，还有很多杂乱的属性值未处理的原因。

心得体会

本次大作业让我学会了如何用 R 语言进行机器学习，这次作业其实和本学期的另一门专业课人工智能的大作业——预测葡萄酒的类型和质量很相似，都属于机器学习，在两次作业的完成中，我逐渐领会了机器学习的常规套路，都是先导入数据，再进行可视化探索数据，在进行数据清晰、特征处理，其中特征处理的部分，我在本次作业中主要是通过观察 summary 和 coefficient 部分的数据，肉眼选择特征，从人工智能大作业的学习中我学到在遇到这种有十几个特征的数据集时其实可以采用特征降维，比如使用 PCA 或者 LDA 的方法，但是由于本次作业和期末月紧联在一起，我没有安排好时间来完善这个部分，以后如果有机会来改进的话我会来做这样的尝试。 本次的大作业学习让我深入了解到了老师上课讲的内容，如第三章的数据准备，教会我如何处理缺失值，如第四章的模型评估，教会我该采用什么系数进行模型评估，如第五章的基本建模方法，教会我如何划分训练集和测试集，如第六章常用建模方法，告诉我该如何建立逻辑回归模型，以及第七章的结果交付，怎么才能漂亮地交一篇代码报告，以往我和同学们都只会用 word 截图+打字说明，现在知道了可以用 knitr 工具包将代码块和注释转成 pdf 文件，不仅提升了美观

性，也增强了可读性。 本次作业将这学期学到的知识都融合起来，并且让我熟悉了 R 语言，使我受益匪浅。感谢老师们本学期辛勤的教学付出！