HW7

xw-zeng

2022-11-27

载入 R 包。

```
library(ggplot2) # plot beautiful graphs
library(rpart) # decision tree
library(rattle) # fancy decision tree visualization
library(pROC) # draw ROC curve
```

定义画 ROC 曲线的函数。

```
show_roc <- function(true, pred, train_or_test){
  roc_curve <- roc(true, pred)
  plot(roc_curve, print.auc = TRUE, auc.polygon = TRUE, legacy.axes = TRUE,
       grid = c(0.1, 0.2), grid.col = c('green', 'red'), max.auc.polygon = TRUE,
       auc.polygon.col = 'skyblue', print.thres = TRUE, main = paste0('ROC Curve on ',
          ifelse(train_or_test == 'train', 'Training', 'Test'), ' Data'))
}
```

征信系列-用户行为数据分析

分析任务 1

读入数据。

```
data <- read.csv('simudata.csv')
```

将变量 black (是否违约) 转化为因子型变量。

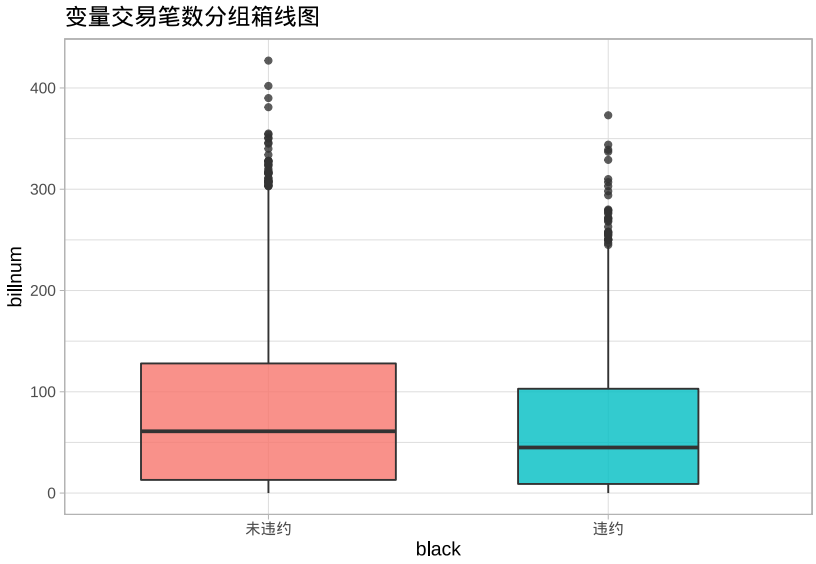
```
data$black <- factor(data$black, levels = c(0, 1), labels = c('未违约', '违约'))
```

仔细阅读说明文档、了解清楚变量含义后，开始进行数据分析任务。

分析任务 2

对变量交易笔数绘制违约组和非违约组的对比箱线图。

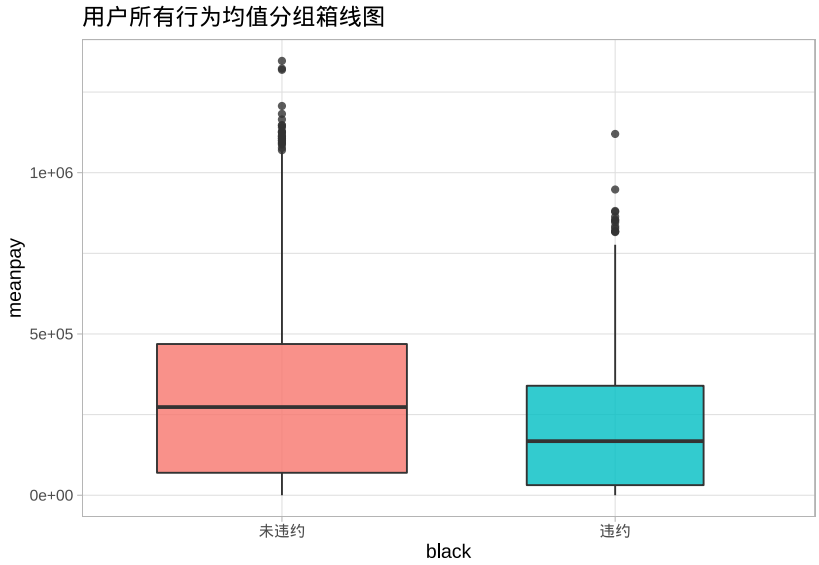
```
ggplot(data, mapping = aes(x = black, y = billnum, fill = black)) +
  geom_boxplot(varwidth = TRUE, alpha = 0.8) +
  labs(title = '变量交易笔数分组箱线图', xlab = '是否违约', ylab = '变量交易笔数') +
  guides(fill = 'none') +
  theme_light()
```



由上图可知，未违约组的样本量大于违约组，未违约组交易笔数的中位数和四分位距都高于违约组。这与我们的常识相符，因为一般来说交易笔数更高的人信用度也更高、更不容易违约，且信用度高的群体交易笔数波动性也较大。

对用户所有行为均值绘制违约组和非违约组的对比箱线图。

```
ggplot(data, mapping = aes(x = black, y = meanpay, fill = black)) +
  geom_boxplot(varwidth = TRUE, alpha = 0.8) +
  labs(title = '用户所有行为均值分组箱线图', xlab = '是否违约', ylab = '用户所有行为均值') +
  guides(fill = 'none') +
  theme_light()
```



由上图可知，未违约组的样本量大于违约组，未违约组用户所有行为均值的中位数和四分位距也都高于违约组。这与我们的常识相符，因为一般来说平均交易金额更高的用户信用度也更高、更不容易违约，且信用度高的群体平均交易金额波动性也较大。

分析任务 3

按照 7:3 的比例划分训练集和测试集。

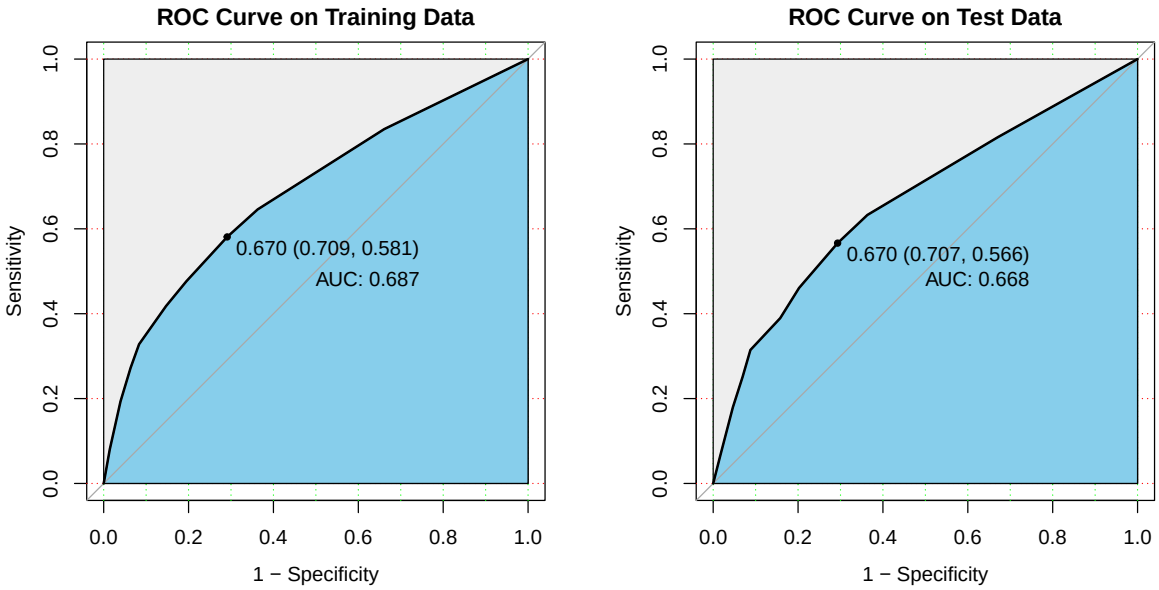
```
set.seed(1234)
idx_train <- sample(1:nrow(data), 0.7 * nrow(data))
data_train <- data[idx_train, ]; data_test <- data[-idx_train, ]
```

用决策树模型在训练集上进行建模，然后使用建立好的模型分别对训练集、测试集进行预测。

```
fit <- rpart(black ~ ., data_train)
pred_train <- predict(fit, data_train, type = 'prob')
pred_test <- predict(fit, data_test, type = 'prob')
```

分别画出模型在训练集、测试集上的 ROC 曲线，AUC 分别为 0.687 和 0.668，模型效果一般。

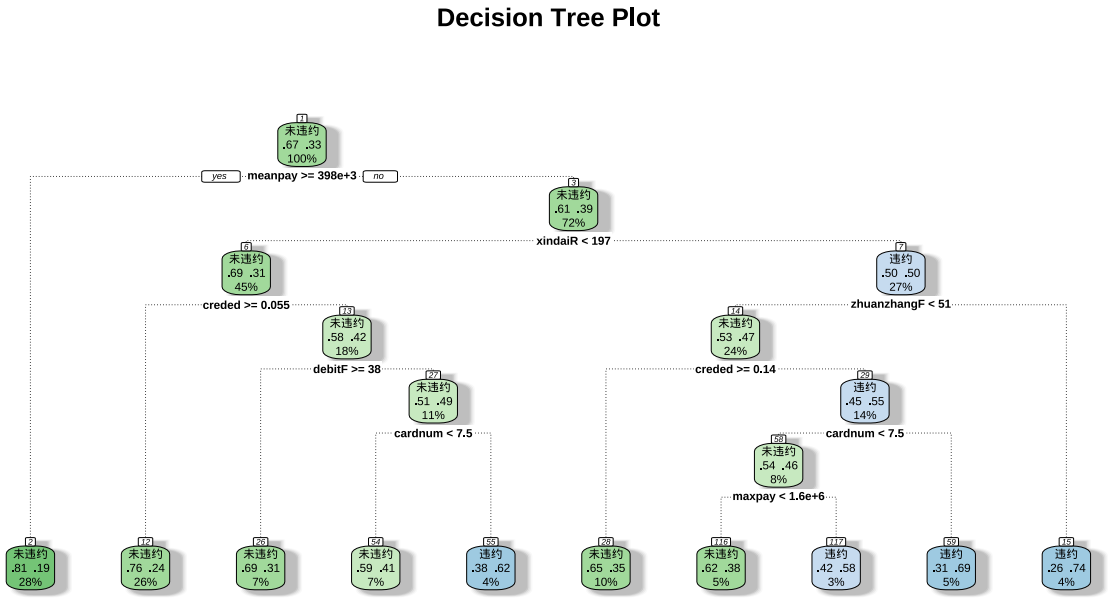
```
par(mfrow=c(1,2))
show_roc(data_train$black, pred_train[, 1], 'train')
show_roc(data_test$black, pred_test[, 1], 'test')
```



分析任务 4

画出第 3 问得到的决策树的图形。

```
fancyRpartPlot(fit, main='Decision Tree Plot', sub='', type = 2)
```



上图已经将决策树的 if-then 规则展示得非常清楚，此处就不再用文字重述了。

由上图可知，对于判断用户是否违约的重要变量有：

- 用户所有交易行为的平均金额 `meanpay`：平均金额越高，用户越不容易违约。
- 用户最近一次小额贷款行为距离数据提取时间的时间间隔 `xindaiR`：上一次小额贷款时间越近，用户越不容易违约。
- 用户借贷比率 `creded`：借贷比率越高，用户越不容易违约。
- 用户银行卡数 `cardnum`：用户银行卡数量越少，用户越不容易违约。

上述解释均符合现实情况。除了这 4 个变量之外，用户使用储蓄卡的频率 `debitF`、用户转账的频率 `zhuanzhangF`、用户所有交易行为的最大金额 `maxpay` 这 3 个变量也出现在了该决策树中，是判断用户是否违约的较重要的变量。

最后查看该决策树中的变量重要性，与上述结果解读相同。

```
fit$variable.importance
```

```
##     meanpay     xindaiR      creded     cardnum zhuanzhangF      debitF
##  83.7654788  66.6307103  65.8036384  33.7034891  25.7649751  16.1865521
##      maxpay      debitM       sidaM     xindaiS zhuanzhangM       sidaF
##   9.6017157   1.0805653   1.0143788   0.9030014   0.5585827   0.4021092
##      jinkaF     xindaiF  zhongxingM   gongjiaoM zhuanzhangR  zhongxingR
##   0.3987074   0.3911866   0.3840156   0.3137898   0.2773827   0.2565922
##     billnum    xiaofeiF      youxiM  zhongxingF         age
##   0.1676514   0.1214648   0.1214648   0.1214648   0.1045966
```

THE END. THANKS! ^_^