```
---
title: "810 Project"
author: "Bo Li U24425931"
date: "2021/2/21"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

install.packages(c("data.table", "ggplot2", "ggthemes", "scales", "rpart", "randomForest",
"glmnet","gbm"))

library(data.table)
library(ggplot2)
library(ggthemes)
library(glmnet)
theme_set(theme_bw())

dd_train <- fread("C:/Users/boli0/Downloads/train.csv")
dd_test <- fread("C:/Users/boli0/Downloads/test.csv")

#Build the model

library(rpart)
fit=rpart(price_range ~.,method = "class", data = dd_train,control = rpart.control(minsplit = 1) , parms =
list(split="information"))
print(fit)

summary(fit)
```

```
> library(rpart)
> fit=rpart(price_range ~.,method = "class", data = dd_train,control = rpart.control(minsplit = 1) , parms = list(split
="information"))
> print(fit)
n= 2000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 2000 1500 0 (0.250000000 0.250000000 0.250000000 0.250000000)
   2) ram< 2235.5 1045  545 0 (0.478468900 0.419138756 0.102392344 0.000000000)
     4) ram< 1106 451   49 0 (0.891352550 0.108647450 0.000000000 0.000000000) *
     5) ram>=1106 594  205 1 (0.164983165 0.654882155 0.180134680 0.000000000)
      10) battery_power< 1108.5 251   96 1 (0.346613546 0.617529880 0.035856574 0.000000000)
        20) ram< 1541 96   21 0 (0.781250000 0.218750000 0.000000000 0.000000000) *
        21) ram>=1541 155   21 1 (0.077419355 0.864516129 0.058064516 0.000000000) *
      11) battery_power>=1108.5 343  109 1 (0.032069971 0.682215743 0.285714286 0.000000000)
        22) ram< 1569.5 157   21 1 (0.063694268 0.866242038 0.070063694 0.000000000) *
        23) ram>=1569.5 186   88 1 (0.005376344 0.526881720 0.467741935 0.000000000)
          46) px_width< 1110 84   11 1 (0.011904762 0.869047619 0.119047619 0.000000000) *
          47) px_width>=1110 102   25 2 (0.000000000 0.245098039 0.754901961 0.000000000) *
   3) ram>=2235.5 955  455 3 (0.000000000 0.064921466 0.411518325 0.523560209)
     6) ram< 3013.5 435  130 2 (0.000000000 0.142528736 0.701149425 0.156321839) *
     7) ram>=3013.5 520   88 3 (0.000000000 0.000000000 0.169230769 0.830769231) *
> #Build the model
>
> library(rpart)
> fit=rpart(price_range ~.,method = "class", data = dd_train,control = rpart.control(minsplit = 1) , parms = list(split
="information"))
> print(fit)
n= 2000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 2000 1500 0 (0.250000000 0.250000000 0.250000000 0.250000000)
   2) ram< 2235.5 1045  545 0 (0.478468900 0.419138756 0.102392344 0.000000000)
     4) ram< 1106 451   49 0 (0.891352550 0.108647450 0.000000000 0.000000000) *
     5) ram>=1106 594  205 1 (0.164983165 0.654882155 0.180134680 0.000000000)
      10) battery_power< 1108.5 251   96 1 (0.346613546 0.617529880 0.035856574 0.000000000)
        20) ram< 1541 96   21 0 (0.781250000 0.218750000 0.000000000 0.000000000) *
        21) ram>=1541 155   21 1 (0.077419355 0.864516129 0.058064516 0.000000000) *
      11) battery_power>=1108.5 343  109 1 (0.032069971 0.682215743 0.285714286 0.000000000)
        22) ram< 1569.5 157   21 1 (0.063694268 0.866242038 0.070063694 0.000000000) *
        23) ram>=1569.5 186   88 1 (0.005376344 0.526881720 0.467741935 0.000000000)
          46) px_width< 1110 84   11 1 (0.011904762 0.869047619 0.119047619 0.000000000) *
          47) px_width>=1110 102   25 2 (0.000000000 0.245098039 0.754901961 0.000000000) *
   3) ram>=2235.5 955  455 3 (0.000000000 0.064921466 0.411518325 0.523560209)
     6) ram< 3013.5 435  130 2 (0.000000000 0.142528736 0.701149425 0.156321839) *
     7) ram>=3013.5 520   88 3 (0.000000000 0.000000000 0.169230769 0.830769231) *
> summary(fit)
Call:
rpart(formula = price_range ~ ., data = dd_train, method = "class",
    parms = list(split = "information"), control = rpart.control(minsplit = 1))
  n= 2000

          CP nsplit rel error    xerror       xstd
1 0.33333333      0 1.0000000 1.0513333 0.01217530
2 0.19400000      1 0.6666667 0.6666667 0.01490712
```

```
> summary(fit)
Call:
rpart(formula = price_range ~ ., data = dd_train, method = "class",
    parms = list(split = "information"), control = rpart.control(minsplit = 1))
  n= 2000

          CP nsplit rel error    xerror       xstd
1 0.33333333      0 1.0000000 1.0513333 0.01217530
2 0.19400000      1 0.6666667 0.6666667 0.01490712
3 0.15800000      2 0.4726667 0.4813333 0.01431950
4 0.01800000      3 0.3146667 0.3300000 0.01286662
5 0.01733333      5 0.2786667 0.3173333 0.01269667
6 0.01000000      7 0.2440000 0.2853333 0.01222762

Variable importance
          ram battery_power     px_height      px_width
           83             6             3             3
         sc_w    int_memory            fc
            1             1             1

Node number 1: 2000 observations,    complexity param=0.3333333
  predicted class=0  expected loss=0.75  P(node) =1
    class counts:   500   500   500   500
   probabilities: 0.250 0.250 0.250 0.250
  left son=2 (1045 obs) right son=3 (955 obs)
  Primary splits:
      ram           < 2235.5 to the left,  improve=937.254700, (0 missing)
      battery_power < 1274   to the left,  improve= 48.124830, (0 missing)
      px_width      < 1645.5 to the left,  improve= 32.232770, (0 missing)
      px_height     < 1258.5 to the left,  improve= 28.001520, (0 missing)
      sc_w          < 10.5   to the left,  improve=  6.734578, (0 missing)
  Surrogate splits:
      sc_w          < 10.5   to the left,  agree=0.537, adj=0.030, (0 split)
      px_height     < 286.5  to the right, agree=0.536, adj=0.028, (0 split)
      battery_power < 648.5  to the right, agree=0.532, adj=0.021, (0 split)
      int_memory    < 60.5   to the left,  agree=0.530, adj=0.016, (0 split)
      fc            < 13.5   to the left,  agree=0.527, adj=0.009, (0 split)

Node number 2: 1045 observations,    complexity param=0.194
  predicted class=0  expected loss=0.5215311  P(node) =0.5225
    class counts:   500   438   107     0
   probabilities: 0.478 0.419 0.102 0.000
  left son=4 (451 obs) right son=5 (594 obs)
  Primary splits:
      ram           < 1106   to the left,  improve=313.640000, (0 missing)
      battery_power < 1463   to the left,  improve= 55.490090, (0 missing)
      px_height     < 642    to the left,  improve= 47.109400, (0 missing)
      px_width      < 1081   to the left,  improve= 34.549000, (0 missing)
      n_cores       < 4.5    to the left,  improve=  6.926756, (0 missing)
  Surrogate splits:
      px_width   < 591.5  to the left,  agree=0.583, adj=0.033, (0 split)
      pc         < 1.5    to the left,  agree=0.574, adj=0.013, (0 split)
      mobile_wt  < 91.5   to the left,  agree=0.572, adj=0.009, (0 split)
      fc         < 17.5   to the right, agree=0.571, adj=0.007, (0 split)
      int_memory < 6.5    to the left,  agree=0.570, adj=0.004, (0 split)

Node number 3: 955 observations,    complexity param=0.158
  predicted class=3  expected loss=0.4764398  P(node) =0.4775
    class counts:     0    62   393   500
   probabilities: 0.000 0.065 0.412 0.524
```

```
Node number 3: 955 observations,    complexity param=0.158
  predicted class=3  expected loss=0.4764398  P(node) =0.4775
    class counts:     0    62   393   500
   probabilities: 0.000 0.065 0.412 0.524
  left son=6 (435 obs) right son=7 (520 obs)
  Primary splits:
      ram           < 3013.5 to the left,  improve=250.343700, (0 missing)
      battery_power < 1353   to the left,  improve= 61.187830, (0 missing)
      px_width      < 1283   to the left,  improve= 45.607970, (0 missing)
      px_height     < 955    to the left,  improve= 38.922220, (0 missing)
      int_memory    < 11.5   to the left,  improve=  7.409328, (0 missing)
  Surrogate splits:
      battery_power < 583    to the left,  agree=0.561, adj=0.037, (0 split)
      int_memory    < 4.5    to the left,  agree=0.554, adj=0.021, (0 split)
      sc_h          < 18.5   to the right, agree=0.551, adj=0.014, (0 split)
      px_height     < 1798.5 to the right, agree=0.549, adj=0.009, (0 split)
      clock_speed   < 2.85   to the right, agree=0.547, adj=0.005, (0 split)

Node number 4: 451 observations
  predicted class=0  expected loss=0.1086475  P(node) =0.2255
    class counts:   402    49     0     0
   probabilities: 0.891 0.109 0.000 0.000

Node number 5: 594 observations,    complexity param=0.018
  predicted class=1  expected loss=0.3451178  P(node) =0.297
    class counts:    98   389   107     0
   probabilities: 0.165 0.655 0.180 0.000
  left son=10 (251 obs) right son=11 (343 obs)
  Primary splits:
      battery_power < 1108.5 to the left,  improve=77.712940, (0 missing)
      ram           < 1508   to the left,  improve=72.278570, (0 missing)
      px_height     < 710.5  to the left,  improve=48.108980, (0 missing)
      px_width      < 1113.5 to the left,  improve=41.963890, (0 missing)
      n_cores       < 4.5    to the left,  improve= 6.147587, (0 missing)
  Surrogate splits:
      px_height  < 116    to the left,  agree=0.603, adj=0.060, (0 split)
      mobile_wt  < 191.5  to the right, agree=0.589, adj=0.028, (0 split)
      int_memory < 2.5    to the left,  agree=0.584, adj=0.016, (0 split)
      ram        < 1972   to the right, agree=0.582, adj=0.012, (0 split)
      pc         < 0.5    to the left,  agree=0.579, adj=0.004, (0 split)

Node number 6: 435 observations
  predicted class=2  expected loss=0.2988506  P(node) =0.2175
    class counts:     0    62   305    68
   probabilities: 0.000 0.143 0.701 0.156

Node number 7: 520 observations
  predicted class=3  expected loss=0.1692308  P(node) =0.26
    class counts:     0     0    88   432
   probabilities: 0.000 0.000 0.169 0.831

Node number 10: 251 observations,    complexity param=0.018
  predicted class=1  expected loss=0.3824701  P(node) =0.1255
    class counts:    87   155     9     0
   probabilities: 0.347 0.618 0.036 0.000
  left son=20 (96 obs) right son=21 (155 obs)
  Primary splits:
      ram       < 1541   to the left,  improve=70.591460, (0 missing)
      px_height < 1026   to the left,  improve=23.187280, (0 missing)
      px_width  < 1158   to the left,  improve=21.725440, (0 missing)
```

```
      ram       < 1541   to the left,  improve=70.591460, (0 missing)
      px_height < 1026   to the left,  improve=23.187280, (0 missing)
      px_width  < 1158   to the left,  improve=21.725440, (0 missing)
      mobile_wt < 139.5  to the right, improve= 5.607004, (0 missing)
      sc_w      < 0.5    to the left,  improve= 5.289705, (0 missing)
  Surrogate splits:
      sc_h        < 18.5   to the right, agree=0.637, adj=0.052, (0 split)
      clock_speed < 2.55   to the right, agree=0.633, adj=0.042, (0 split)
      m_dep       < 0.15   to the left,  agree=0.633, adj=0.042, (0 split)
      px_height   < 76     to the left,  agree=0.629, adj=0.031, (0 split)
      px_width    < 577    to the left,  agree=0.629, adj=0.031, (0 split)

Node number 11: 343 observations,    complexity param=0.01733333
  predicted class=1  expected loss=0.3177843  P(node) =0.1715
    class counts:    11   234    98     0
   probabilities: 0.032 0.682 0.286 0.000
  left son=22 (157 obs) right son=23 (186 obs)
  Primary splits:
      ram           < 1569.5 to the left,  improve=39.657850, (0 missing)
      px_width      < 1112.5 to the left,  improve=34.164960, (0 missing)
      px_height     < 698.5  to the left,  improve=33.426160, (0 missing)
      battery_power < 1466.5 to the left,  improve=14.700580, (0 missing)
      n_cores       < 4.5    to the left,  improve= 6.175137, (0 missing)
  Surrogate splits:
      three_g       < 0.5    to the left,  agree=0.589, adj=0.102, (0 split)
      px_width      < 1779   to the right, agree=0.580, adj=0.083, (0 split)
      talk_time     < 15.5   to the right, agree=0.574, adj=0.070, (0 split)
      battery_power < 1641.5 to the right, agree=0.571, adj=0.064, (0 split)
      px_height     < 1443.5 to the right, agree=0.571, adj=0.064, (0 split)

Node number 20: 96 observations
  predicted class=0  expected loss=0.21875  P(node) =0.048
    class counts:    75    21     0     0
   probabilities: 0.781 0.219 0.000 0.000

Node number 21: 155 observations
  predicted class=1  expected loss=0.1354839  P(node) =0.0775
    class counts:    12   134     9     0
   probabilities: 0.077 0.865 0.058 0.000

Node number 22: 157 observations
  predicted class=1  expected loss=0.133758  P(node) =0.0785
    class counts:    10   136    11     0
   probabilities: 0.064 0.866 0.070 0.000

Node number 23: 186 observations,    complexity param=0.01733333
  predicted class=1  expected loss=0.4731183  P(node) =0.093
    class counts:     1    98    87     0
   probabilities: 0.005 0.527 0.468 0.000
  left son=46 (84 obs) right son=47 (102 obs)
  Primary splits:
      px_width      < 1110   to the left,  improve=41.366530, (0 missing)
      px_height     < 708    to the left,  improve=30.661050, (0 missing)
      battery_power < 1484   to the left,  improve=17.788700, (0 missing)
      ram           < 1896.5 to the left,  improve=12.831950, (0 missing)
      n_cores       < 4.5    to the left,  improve= 4.447158, (0 missing)
  Surrogate splits:
      px_height     < 708    to the left,  agree=0.763, adj=0.476, (0 split)
      battery_power < 1488   to the left,  agree=0.624, adj=0.167, (0 split)
      n_cores       < 2.5    to the left,  agree=0.608, adj=0.131, (0 split)
      dual_sim      < 0.5    to the right, agree=0.591, adj=0.095, (0 split)
```

```
  predicted class=1  expected loss=0.3177843  P(node) =0.1715
    class counts:    11   234    98     0
   probabilities: 0.032 0.682 0.286 0.000
  left son=22 (157 obs) right son=23 (186 obs)
  Primary splits:
      ram           < 1569.5 to the left,  improve=39.657850, (0 missing)
      px_width      < 1112.5 to the left,  improve=34.164960, (0 missing)
      px_height     < 698.5  to the left,  improve=33.426160, (0 missing)
      battery_power < 1466.5 to the left,  improve=14.700580, (0 missing)
      n_cores       < 4.5    to the left,  improve= 6.175137, (0 missing)
  Surrogate splits:
      three_g       < 0.5    to the left,  agree=0.589, adj=0.102, (0 split)
      px_width      < 1779   to the right, agree=0.580, adj=0.083, (0 split)
      talk_time     < 15.5   to the right, agree=0.574, adj=0.070, (0 split)
      battery_power < 1641.5 to the right, agree=0.571, adj=0.064, (0 split)
      px_height     < 1443.5 to the right, agree=0.571, adj=0.064, (0 split)

Node number 20: 96 observations
  predicted class=0  expected loss=0.21875  P(node) =0.048
    class counts:    75    21     0     0
   probabilities: 0.781 0.219 0.000 0.000

Node number 21: 155 observations
  predicted class=1  expected loss=0.1354839  P(node) =0.0775
    class counts:    12   134     9     0
   probabilities: 0.077 0.865 0.058 0.000

Node number 22: 157 observations
  predicted class=1  expected loss=0.133758  P(node) =0.0785
    class counts:    10   136    11     0
   probabilities: 0.064 0.866 0.070 0.000

Node number 23: 186 observations,    complexity param=0.01733333
  predicted class=1  expected loss=0.4731183  P(node) =0.093
    class counts:     1    98    87     0
   probabilities: 0.005 0.527 0.468 0.000
  left son=46 (84 obs) right son=47 (102 obs)
  Primary splits:
      px_width      < 1110   to the left,  improve=41.366530, (0 missing)
      px_height     < 708    to the left,  improve=30.661050, (0 missing)
      battery_power < 1484   to the left,  improve=17.788700, (0 missing)
      ram           < 1896.5 to the left,  improve=12.831950, (0 missing)
      n_cores       < 4.5    to the left,  improve= 4.447158, (0 missing)
  Surrogate splits:
      px_height     < 708    to the left,  agree=0.763, adj=0.476, (0 split)
      battery_power < 1488   to the left,  agree=0.624, adj=0.167, (0 split)
      n_cores       < 2.5    to the left,  agree=0.608, adj=0.131, (0 split)
      dual_sim      < 0.5    to the right, agree=0.591, adj=0.095, (0 split)
      pc            < 18.5   to the right, agree=0.586, adj=0.083, (0 split)

Node number 46: 84 observations
  predicted class=1  expected loss=0.1309524  P(node) =0.042
    class counts:     1    73    10     0
   probabilities: 0.012 0.869 0.119 0.000

Node number 47: 102 observations
  predicted class=2  expected loss=0.245098  P(node) =0.051
    class counts:     0    25    77     0
   probabilities: 0.000 0.245 0.755 0.000
```