DM_DecisionTree.R

mac

2019-11-08

```
Airb=read.csv("~/Documents/R/AB_NYC_2019.csv",header=T,na.strings="?")

attach(Airb)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X1
##   Referenced from: /Library/Frameworks/R.framework/Versions/3.3/Resources/modules/R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
```

```
## Loading required package: RSQLite
```

```
library(rpart)

price_x=Airb$price>100
Airb=data.frame(Airb,price_x)

TRain=sqldf("select longitude,neighbourhood_group,latitude,price,minimum_nights,
            availability_365,price_x from Airb where price!=0 and
            price<1000 and id%10!=1")


TEst=sqldf("select longitude,neighbourhood_group,latitude,price,minimum_nights,
            availability_365,price_x from Airb where price!=0 and
           price<1000 and id%10==1")

# grow tree
fit = rpart(price_x~longitude+latitude+minimum_nights+
            availability_365,method="class", data=TRain)


summary(fit) # detailed summary of splits
```

```
## Call:
## rpart(formula = price_x ~ longitude + latitude + minimum_nights +
##     availability_365, data = TRain, method = "class")
##   n= 43709
##
##           CP nsplit rel error    xerror        xstd
## 1 0.33303870      0 1.0000000 1.0000000 0.004862273
## 2 0.03163379      1 0.6669613 0.6684034 0.004568699
## 3 0.01789480      2 0.6353275 0.6365836 0.004510295
## 4 0.01000000      5 0.5816431 0.5855973 0.004404149
##
```

```
## Variable importance
##        longitude         latitude   minimum_nights availability_365
##               71               25                3                1
##
## Node number 1: 43709 observations,    complexity param=0.3330387
##   predicted class=TRUE   expected loss=0.491798  P(node) =1
##     class counts: 21496 22213
##    probabilities: 0.492 0.508
##   left son=2 (25987 obs) right son=3 (17722 obs)
##   Primary splits:
##       longitude        < -73.96338 to the right, improve=2730.33700, (0 missing)
##       latitude         < 40.70557  to the left,  improve= 888.07040, (0 missing)
##       minimum_nights   < 1.5       to the left,  improve= 422.24230, (0 missing)
##       availability_365 < 100.5     to the left,  improve=  94.01156, (0 missing)
##   Surrogate splits:
##       minimum_nights < 28.5      to the left,  agree=0.607, adj=0.031, (0 split)
##       latitude       < 40.6504   to the right, agree=0.599, adj=0.010, (0 split)
##
## Node number 2: 25987 observations,    complexity param=0.0178948
##   predicted class=FALSE  expected loss=0.3622581  P(node) =0.5945457
##     class counts: 16573  9414
##    probabilities: 0.638 0.362
##   left son=4 (13214 obs) right son=5 (12773 obs)
##   Primary splits:
##       longitude        < -73.94136 to the right, improve=465.7822, (0 missing)
##       latitude         < 40.70794  to the left,  improve=192.6123, (0 missing)
##       minimum_nights   < 1.5       to the left,  improve=176.0664, (0 missing)
##       availability_365 < 2.5       to the left,  improve= 61.0164, (0 missing)
##   Surrogate splits:
##       availability_365 < 22.5      to the right, agree=0.574, adj=0.133, (0 split)
##       latitude         < 40.70741  to the left,  agree=0.573, adj=0.131, (0 split)
##       minimum_nights   < 2.5       to the left,  agree=0.549, adj=0.082, (0 split)
##
## Node number 3: 17722 observations,    complexity param=0.03163379
##   predicted class=TRUE   expected loss=0.2777903  P(node) =0.4054543
##     class counts:  4923 12799
##    probabilities: 0.278 0.722
##   left son=6 (1566 obs) right son=7 (16156 obs)
##   Primary splits:
##       latitude         < 40.66046  to the left,  improve=663.08500, (0 missing)
##       longitude        < -74.01775 to the left,  improve=209.16150, (0 missing)
##       minimum_nights   < 1.5       to the left,  improve=150.30410, (0 missing)
##       availability_365 < 102.5     to the left,  improve= 42.87873, (0 missing)
##   Surrogate splits:
##       longitude < -74.0178  to the left,  agree=0.941, adj=0.337, (0 split)
##
## Node number 4: 13214 observations
##   predicted class=FALSE  expected loss=0.2691842  P(node) =0.3023176
##     class counts:  9657  3557
##    probabilities: 0.731 0.269
##
## Node number 5: 12773 observations,    complexity param=0.0178948
##   predicted class=FALSE  expected loss=0.4585454  P(node) =0.2922281
##     class counts:  6916  5857
##    probabilities: 0.541 0.459 
##   left son=10 (4837 obs) right son=11 (7936 obs)
##   Primary splits:
##       latitude         < 40.70968  to the left,  improve=234.79810, (0 missing)
##       availability_365 < 2.5       to the left,  improve= 78.80227, (0 missing)
##       minimum_nights   < 1.5       to the left,  improve= 76.98484, (0 missing)
##       longitude        < -73.94787 to the right, improve= 29.56019, (0 missing)
## 
## Node number 6: 1566 observations
##   predicted class=FALSE  expected loss=0.2828863  P(node) =0.03582786
##     class counts:  1123   443
##    probabilities: 0.717 0.283 
## 
## Node number 7: 16156 observations
##   predicted class=TRUE   expected loss=0.2352067  P(node) =0.3696264
##     class counts:  3800 12356
##    probabilities: 0.235 0.765 
## 
## Node number 10: 4837 observations
##   predicted class=FALSE  expected loss=0.3357453  P(node) =0.1106637
##     class counts:  3213  1624
##    probabilities: 0.664 0.336 
## 
## Node number 11: 7936 observations,    complexity param=0.0178948
##   predicted class=TRUE   expected loss=0.4666079  P(node) =0.1815644
##     class counts:  3703  4233
##    probabilities: 0.467 0.533 
##   left son=22 (1584 obs) right son=23 (6352 obs)
##   Primary splits:
##       latitude         < 40.81141  to the right, improve=210.03780, (0 missing)
##       minimum_nights   < 1.5       to the left,  improve= 58.39440, (0 missing)
##       availability_365 < 2.5       to the left,  improve= 48.67050, (0 missing)
##       longitude        < -73.94632 to the right, improve= 45.30843, (0 missing)
## 
## Node number 22: 1584 observations
##   predicted class=FALSE  expected loss=0.3030303  P(node) =0.03623968
##     class counts:  1104   480
##    probabilities: 0.697 0.303 
## 
## Node number 23: 6352 observations
##   predicted class=TRUE   expected loss=0.4091625  P(node) =0.1453248
##     class counts:  2599  3753
##    probabilities: 0.409 0.591
```

```
#rpart.plot(fit) # rpart.plot is not available

predtree=predict(fit,newdata=TEst,type="class")
#result
table(TEst$price_x,predtree)
```

```
##        predtree
##         FALSE TRUE
##   FALSE  1701  720
##   TRUE    659 1797
```

```r
cm = as.matrix(table(Actual = TEst$price_x, Predicted = predtree))
accu=sum(diag(cm))/length(TEst$price_x)
#accuracy
message(accu)
```

```
## 0.717244207504613
```