ScriptProjeto1.R

braulioaraujo

2020-10-28

```
## Ao avaliar os dados, decidi não incluir a variável attributed_time pois ela só consta
## nos casos de realização de downloads. Tentei alguns algoritimos e o Random Forest se
## mostrou melhor.

library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(randomForest)
```

```
## randomForest 4.6-14
```

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

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

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

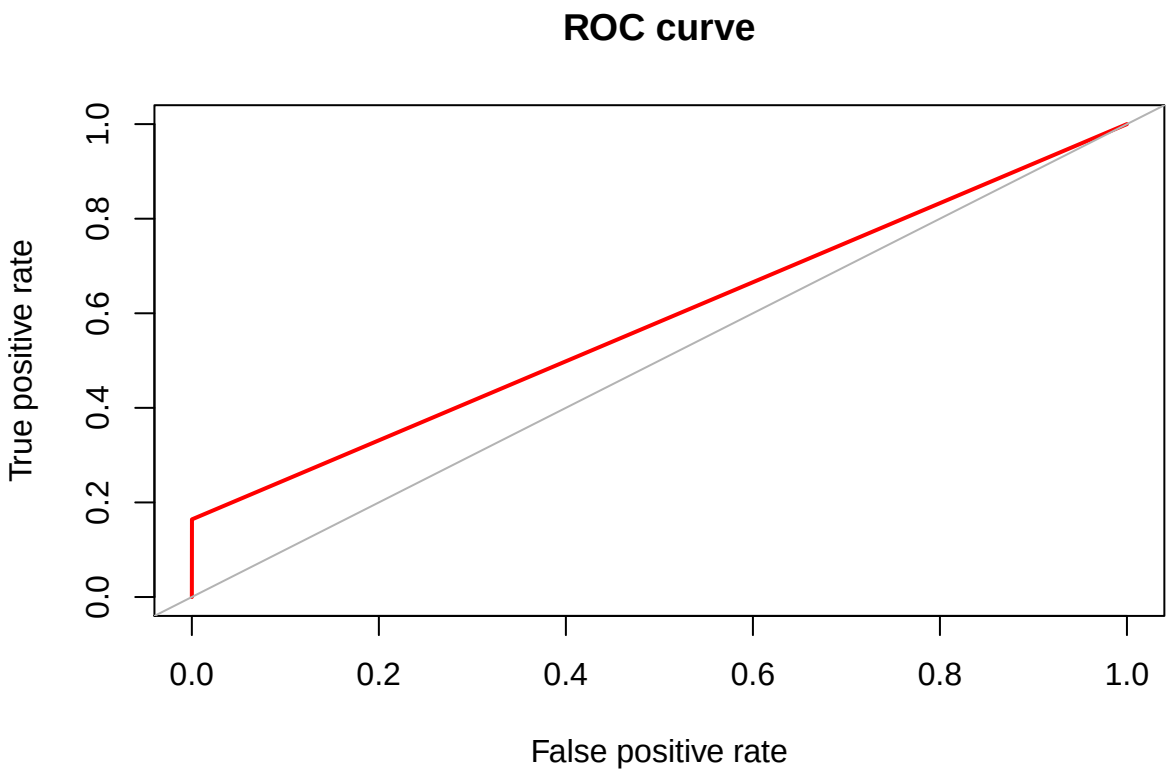
```
library(ROSE)
```

```
## Loaded ROSE 0.0-3
```

```r
dt = read.csv(file='train_sample.csv')

# Ajuste nas variáveis

dt$click_time = ymd_hms(dt$click_time)
dt$attributed_time = ymd_hms(dt$attributed_time)
dt$is_attributed = as.factor(dt$is_attributed)

# divisao de treino e teste

amostra = sample(2,100000,replace=T, prob=c(0.7,0.3))
treino = dt[amostra==1,]
teste = dt[amostra==2,]

# Construindo o modelo

modelo = randomForest(is_attributed ~ ip + app + device + os + channel + click_time, treino)

# Fazendo as previsões

previsoes = predict(modelo, teste)

# Avaliando o modelo

confusionMatrix(teste$is_attributed, previsoes)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 29936     4
##          1    56    11
##
##                Accuracy : 0.998
##                  95% CI : (0.9974, 0.9985)
##     No Information Rate : 0.9995
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2677
##
##  Mcnemar's Test P-Value : 4.577e-11
##
##             Sensitivity : 0.9981
##             Specificity : 0.7333
##          Pos Pred Value : 0.9999
##          Neg Pred Value : 0.1642
##              Prevalence : 0.9995
##          Detection Rate : 0.9976
##    Detection Prevalence : 0.9978
##       Balanced Accuracy : 0.8657
##
##        'Positive' Class : 0
##
```

```r
roc.curve(teste$is_attributed, previsoes, plotit = T, col = "red")
```




```
## Area under the curve (AUC): 0.582
```