lista6

Klysman Rezende e Matheus Cougias

09/09/2020

Leitura dos dados

Realiza a leitura do banco de dados corretamente, além de renomear as colunas de acordo com o especificado na fonte.

```
## Loading required package: readr
```

```
## Loading required package: pROC
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
## Loading required package: rpart
```

```
## Loading required package: rpart.plot
```

```
## Loading required package: randomForest
```

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

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

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

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

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

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

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_character(),
##   X3 = col_double(),
##   X4 = col_character(),
##   X5 = col_double(),
##   X6 = col_character(),
##   X7 = col_character(),
##   X8 = col_character(),
##   X9 = col_character(),
##   X10 = col_character(),
##   X11 = col_double(),
##   X12 = col_double(),
##   X13 = col_double(),
##   X14 = col_character(),
##   X15 = col_character()
## )
```

Análise Exploratória

Através de uma análise básica da relação entre a variável de renda e as demais variáveis, percebe-se que um perfil "otimista" de pessoa que tenha renda acima de 50 mil pode ser dado por: homem branco norte-americano casado, com alto nível de escolaridade e idade entre 40 e 50 anos, sendo dono de sua própria empresa, além de trabalhar entre 40 e 50 horas semanais. Por questão de economia de espaço no relatório, os gráficos estão em comentário. Além disso, temos que 24.08% das pessoas amostradas no banco de dados representam aquelas com renda acima de 50 mil.

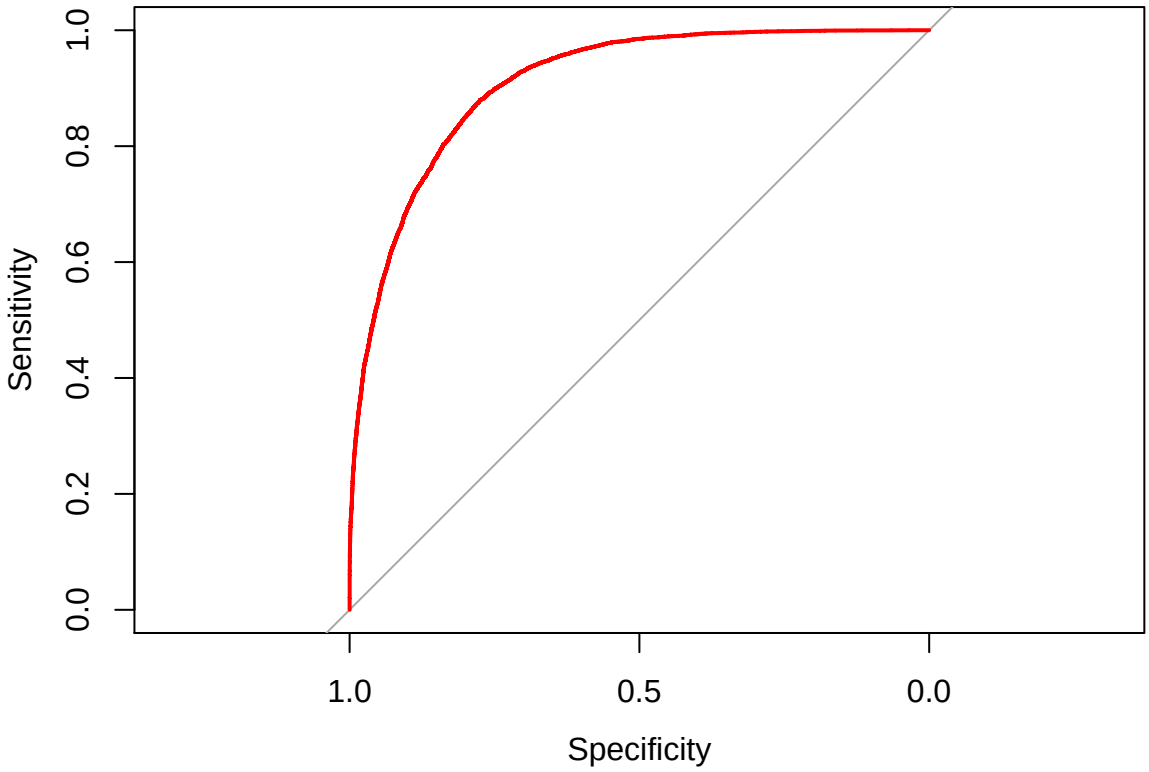
```
##
##    <=50K     >50K
## 0.759183 0.240817
```

Modelo Logístico

A função step realizou o corte de somente uma variável do problema (Num_escolaridade), mostrando que as variáveis não possuem um grau de relacionamento muito grande. Como analizado anteriormente, dentro de cada variável pode-se observar qual "padrão" gerará um melhor valor de modo que a probabilidade da renda do indivíduo ser acima de 50 mil seja maior. As características retiradas da análise exploratória para que o indivíduo possua maior probabilidade de renda acima de 50 mil é similar ao retirado do resultado da regressão logística. A área sobre a curva ROC utilizando a regressão logística é de 90.89%, podendo ser considerado como alto.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```




```
## Area under the curve: 0.9089
```

Validação Modelo Logístico

Utilizando a validação cruzada do modelo de regressão logística, a área sobre a curva ROC é de 90.30%, apresentando um resultado alto e extremamente próximo do da própria regressão.

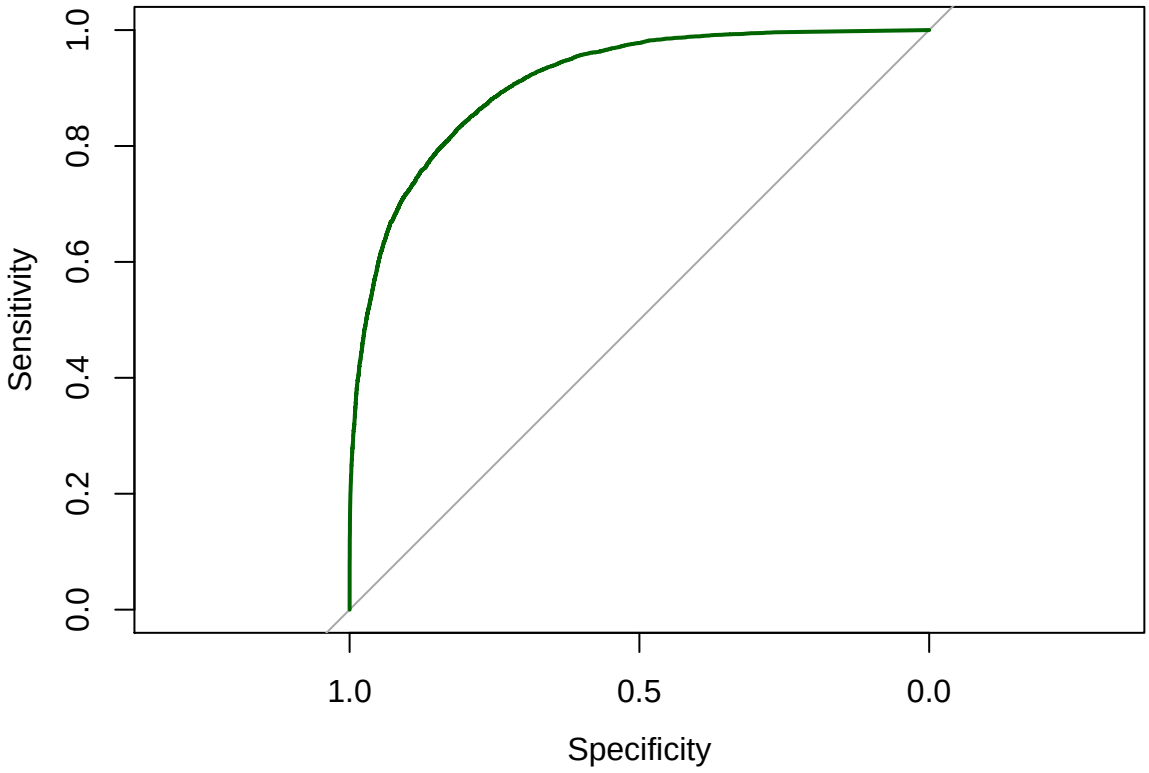
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Setting levels: control = <=50K, case = >50K

## Setting direction: controls < cases

## Area under the curve: 0.9023
```

Random Forest

A partir da utilização do modelo de Random Forest, a área sobre a curva ROC teve um pequeno aumento para 91.05%.

```
## Setting levels: control = <=50K, case = >50K
```

```
## Setting direction: controls < cases
```




```
## Area under the curve: 0.9105
```

Validação Random Forest

Assim, a validação da Random Forest objetve um resultado de 91.04% de área abaixo da curva ROC.

```
## Setting levels: control = <=50K, case = >50K
```

```
## Setting direction: controls > cases
```

```
## Area under the curve: 0.9057
```