20210327-manha-exercicio-titanic.R

rstudio-user

2021-04-10

```
library(titanic)

# Define os subconjuntos
train <- titanic_train
test <- titanic_test
test <- merge(test, titanic_gender_class_model, by="PassengerId")

# Verificando as variáveis
str(train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```
# Verificando se há dados ausentes
colSums(is.na(train))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

```
colSums(is.na(test))
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0          86           0
##       Parch      Ticket        Fare       Cabin    Embarked    Survived
##           0           0           1           0           0           0
```

```
# Verifica se há valores vazios
colSums(train == '')
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

```
colSums(test == '')
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0          NA           0
##       Parch      Ticket        Fare       Cabin    Embarked    Survived
##           0           0          NA         327           0           0
```

```
# Remover valores faltantes e vazios
train <- train[-which(train$Embarked == ""),]
test <- test[-which(is.na(test$Fare)),]

# Colocando a mediana para valores faltantes
train$Age[is.na(train$Age)] <- median(train$Age, na.rm=T)
test$Age[is.na(test$Age)] <- median(test$Age, na.rm=T)

# Remover variáveis não necessárias
train <- subset(train, select = -c(Cabin, PassengerId, Ticket, Name))
test <- subset(test, select = -c(Cabin, PassengerId, Ticket, Name))

# Converter colunas para fatores
for (i in c("Survived","Pclass","Sex","Embarked")){
  train[,i] <- as.factor(train[,i])
}
for (j in c("Survived","Pclass","Sex","Embarked")){
  test[,j] <- as.factor(test[,j])
}

# Correlação das variáveis
library(dlookr)
```

```
## Either Arial Narrow or Liberation Sans Narrow fonts are required to Viz.
## Please use dlookr::import_liberation() to install Liberation Sans Narrow font.
```

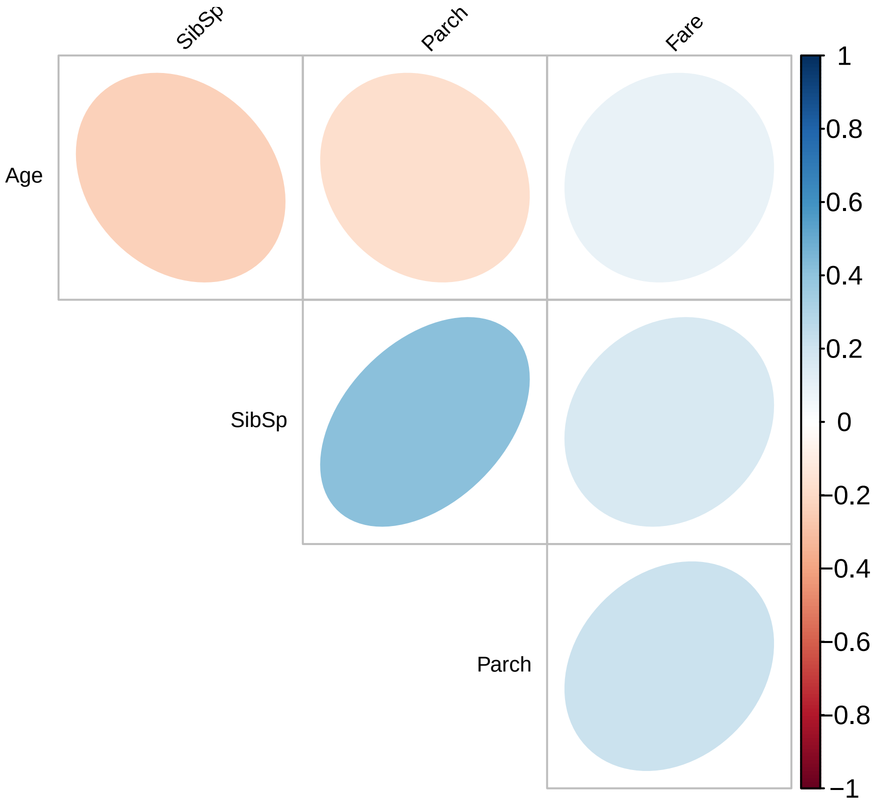
```
##
## Attaching package: 'dlookr'
```

```
## The following object is masked from 'package:base':
##
##     transform
```

```
correlate(train)
```

```
## # A tibble: 12 x 3
##    var1  var2  coef_corr
##    <fct> <fct>     <dbl>
##  1 SibSp Age     -0.233
##  2 Parch Age     -0.171
##  3 Fare  Age      0.0937
##  4 Age   SibSp   -0.233
##  5 Parch SibSp    0.415
##  6 Fare  SibSp    0.161
##  7 Age   Parch   -0.171
##  8 SibSp Parch    0.415
##  9 Fare  Parch    0.218
## 10 Age   Fare     0.0937
## 11 SibSp Fare      0.161
## 12 Parch Fare      0.218
```

```
plot_correlate(train)
```



```
# Removendo linhas com dados ausentes
train <- train[complete.cases(train),]

# Vendo se a classe está balanceada
table(train$Survived)
```

```
##
##   0   1
## 549 340
```

```
prop.table(table(train$Survived))
```

```
##
##         0         1
## 0.6175478 0.3824522
```

```
# Modelo 1
mod1 <- glm(formula = Survived ~ ., data = train, family = "binomial")
mod1
```

```
##
## Call:  glm(formula = Survived ~ ., family = "binomial", data = train)
##
## Coefficients:
## (Intercept)      Pclass2      Pclass3      Sexmale          Age        SibSp
##    4.062486    -0.911903    -2.144097    -2.710309    -0.038752    -0.320495
##       Parch         Fare    EmbarkedQ    EmbarkedS  
##   -0.091313     0.002304    -0.057728    -0.440140  
## 
## Degrees of Freedom: 888 Total (i.e. Null);  879 Residual
## Null Deviance:       1183 
## Residual Deviance: 784.4     AIC: 804.4
```

```
summary(mod1)
```

```
## 
## Call:
## glm(formula = Survived ~ ., family = "binomial", data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.6169  -0.6094  -0.4191   0.6126   2.4527  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  4.062486   0.472734   8.594  < 2e-16 ***
## Pclass2     -0.911903   0.297391  -3.066  0.00217 ** 
## Pclass3     -2.144097   0.297668  -7.203 5.89e-13 ***
## Sexmale     -2.710309   0.201224 -13.469  < 2e-16 ***
## Age         -0.038752   0.007873  -4.922 8.55e-07 ***
## SibSp       -0.320495   0.109056  -2.939  0.00329 ** 
## Parch       -0.091313   0.118850  -0.768  0.44231    
## Fare         0.002304   0.002462   0.936  0.34940    
## EmbarkedQ   -0.057728   0.381060  -0.151  0.87959    
## EmbarkedS   -0.440140   0.239533  -1.837  0.06614 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  784.42  on 879  degrees of freedom
## AIC: 804.42
## 
## Number of Fisher Scoring iterations: 5
```

```
# É possível identificar as variáveis significantes: Pclass2, Pclass3, Sexmale, Age e SibSp

exp(mod1$coefficients)
```

```
## (Intercept)     Pclass2     Pclass3     Sexmale         Age       SibSp 
## 58.11859699  0.40175907  0.11717379  0.06651623  0.96198937  0.72578953 
##       Parch        Fare   EmbarkedQ   EmbarkedS 
##  0.91273242  1.00230689  0.94390705  0.64394615
```

```
# Algumas análises:
# A cada sobrevivente, 0.40 pessoas da segunda classe sobreviveram
# A cada sobrevivente, 0.11 pessoas da terceira classe sobreviveram
# A cada sobrevivente, 0.066 homens sobreviveram
```