Trabalho de R Modelos - FIAP

Carregando as bibliotecas

```
#install.packages("ggcorrplot")
#install.packages("rattle")

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(ggcorrplot)
```

Carregando os dados

```
ds <- read_delim("consolidado_para_analise.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE, show_col_types = FALSE)
```

- Analisando quantidade de observações e variáveis

```
dim(ds)
```

```
## [1] 473  18
```

Conhecendo as variáveis presentes neste dataset

```
names(ds)
```

```
##  [1] "ID"                  "DataNascimento"      "Sexo"
##  [4] "TempodeServiço"      "EstadoCivil"         "NumerodeFilhos"
##  [7] "TempodeResidencia"   "Conta"               "salario"
## [10] "data_atual"          "faixa_salario"       "default"
## [13] "default1"            "QtdaParcelas"        "Atraso"
## [16] "ValorEmprestimo"     "QtdaPagas"           "comprometido_de_renda"
```

- Anotações importantes para análise dos dados:

```
# ID -> Identificador

# Qualitativas:
# Sexo -> nominal (Feminino, Masculino)
# EstadoCivil -> nominal (1,2,3,4)
# Conta -> nominal (empresa, Particular)
# faixa_salario -> nominal (A,B,C,D)
# Atraso -> nominal (Sim, Não)

# Quantitativas:
# NumerodeFilhos -> discreta
# TempodeServiço -> discreta
# TempodeResidencia -> discreta
# salario -> contínua
# QtdaParcelas -> discreta
# ValorEmprestimo -> contínua
# QtdaPagas -> discreta
# comprometido_de_renda -> contínua

# Excluir:
# data_atual
# ID
# DataNascimento

# Preditoras:
# default -> Classificadora
# default1 -> Regressora
```

- Criando um novo dataset para retirar algumas colunas

```
ds_analise <- select(ds, -(ID), -(data_atual), -(DataNascimento))
ds_analise
```

Sexo <chr>	TempodeServiço <dbl>	EstadoCivil <dbl>	NumerodeFilhos <dbl>	TempodeResidencia <dbl>	Conta <chr>
Masculino	98	3	0	7	Particular
Masculino	98	3	0	43	Particular
Feminino	98	1	0	7	empresa
Feminino	98	1	0	13	empresa
Masculino	98	2	3	8	Particular
Masculino	98	2	0	26	Particular
Masculino	98	3	0	40	Particular
Feminino	98	3	0	15	empresa
Feminino	98	2	6	34	empresa
Feminino	98	1	0	0	empresa

1-10 of 473 rows | 1-7 of 15 columns Previous **1** 2 3 4 5 6 … 48 Next

- Verificando as variavéis que ficou no dataset

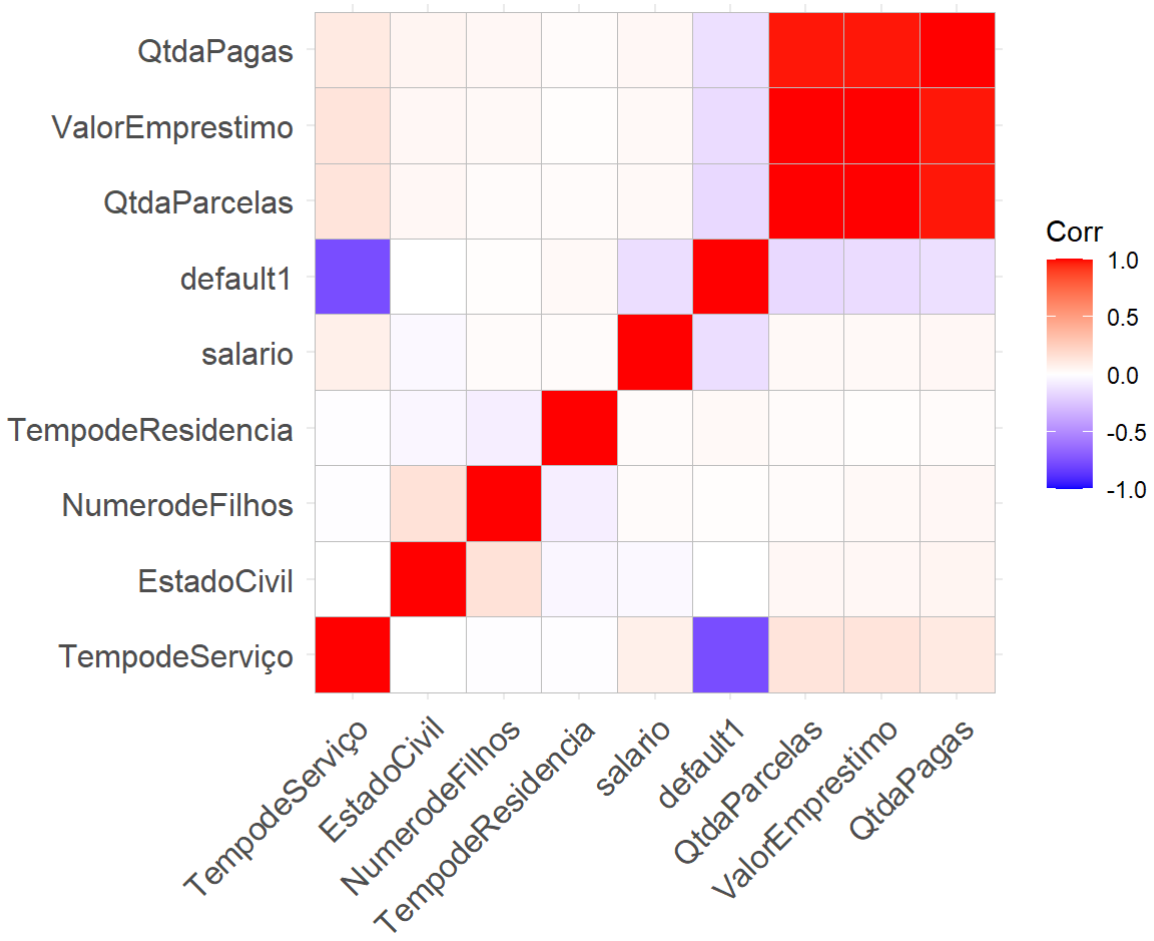
```
names(ds_analise)
```

```
##  [1] "Sexo"                 "TempodeServiço"       "EstadoCivil"
##  [4] "NumerodeFilhos"       "TempodeResidencia"    "Conta"
##  [7] "salario"              "faixa_salario"        "default"
## [10] "default1"             "QtdaParcelas"         "Atraso"
## [13] "ValorEmprestimo"      "QtdaPagas"            "comprometido_de_renda"
```

- Verificando a correlação das variáveis númericas

```
# NumerodeFilhos -> discreta
# TempodeServiço -> discreta
# TempodeResidencia -> discreta
# salario -> contínua
# QtdaParcelas -> discreta
# ValorEmprestimo -> contínua
# QtdaPagas -> discreta
# comprometido_de_renda -> contínua

ds_numericos <- select_if(ds_analise, is.numeric)
correl <-cor(ds_numericos)
ggcorrplot(correl)
```



```
# As variáveis -> QtdaPagas, ValorEmprestimo e QtdaParcelas estão muito correlacionadas,
# irei verificar como o modelo se comporta em relação a elas, e então irei decidir se tiro al
guma
# menos importante.
```

A. Criar o modelo de regressão logística com a variável target DEFAULT1 e interpretar os coeficientes e verificar o quanto o modelo teve de acurácia.

```
logistica <- glm(default1 ~ Sexo + EstadoCivil + NumerodeFilhos + TempodeResidencia +
                   faixa_salario + QtdaParcelas + Atraso + ValorEmprestimo +
                  QtdaPagas + Conta + salario, family = binomial, data = ds_analise)
summary(logistica)
```

```
##
## Call:
## glm(formula = default1 ~ Sexo + EstadoCivil + NumerodeFilhos +
##     TempodeResidencia + faixa_salario + QtdaParcelas + Atraso +
##     ValorEmprestimo + QtdaPagas + Conta + salario, family = binomial,
##     data = ds_analise)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1001  -1.0018   0.6065   0.7930   1.4440
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.676e+00  7.680e-01   3.484 0.000493 ***
## SexoMasculino     -1.957e-01  2.598e-01  -0.753 0.451376
## EstadoCivil        1.368e-02  1.095e-01   0.125 0.900562
## NumerodeFilhos     1.481e-02  6.101e-02   0.243 0.808170
## TempodeResidencia  1.117e-02  1.145e-02   0.975 0.329453
## faixa_salarioB    -5.794e-01  3.309e-01  -1.751 0.079967 .
## faixa_salarioC    -3.540e-02  7.756e-01  -0.046 0.963597
## faixa_salarioD    -7.988e-01  1.285e+00  -0.622 0.534101
## QtdaParcelas      -2.847e+00  9.141e+01  -0.031 0.975154
## AtrasoSim          5.579e-01  3.290e-01   1.696 0.089876 .
## ValorEmprestimo    6.689e-02  2.612e+00   0.026 0.979568
## QtdaPagas          1.339e+00  3.990e-01   3.355 0.000792 ***
## ContaParticular           NA         NA      NA       NA
## salario           -7.637e-06  1.923e-05  -0.397 0.691232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 556.26  on 472  degrees of freedom
## Residual deviance: 511.20  on 460  degrees of freedom
## AIC: 537.2
##
## Number of Fisher Scoring iterations: 12
```

```
# * Através desta  primeira analise, a variável QtdaPagas apresentou mais significante que as demais
# ValorEmprestimo e QtdaParcelas, ainda analisando o caso de correlação
# Conta apresentou valores NA (1 not defined because of singularities), já vou excluir
```

- Retirando a variável QtdaParcelas, pois creio que o valoremprestimo seja mais relevante

```
logistica <- glm(default1 ~ Sexo +
                           EstadoCivil +
                           NumerodeFilhos +
                           TempodeResidencia +
                           faixa_salario  +
                           Atraso +
                           ValorEmprestimo +
                           QtdaPagas+
                           salario, family = binomial, data = ds_analise)
summary(logistica)
```

```
##
## Call:
## glm(formula = default1 ~ Sexo + EstadoCivil + NumerodeFilhos +
##     TempodeResidencia + faixa_salario + Atraso + ValorEmprestimo +
##     QtdaPagas + salario, family = binomial, data = ds_analise)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2220  -1.0240   0.6251   0.7935   1.4789
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.822e+00  7.643e-01   3.692 0.000222 ***
## SexoMasculino     -1.940e-01  2.584e-01  -0.751 0.452765
## EstadoCivil        3.534e-03  1.086e-01   0.033 0.974033
## NumerodeFilhos     2.233e-02  6.073e-02   0.368 0.713074
## TempodeResidencia  7.504e-03  1.122e-02   0.669 0.503680
## faixa_salarioB    -5.852e-01  3.292e-01  -1.778 0.075438 .
## faixa_salarioC     5.033e-02  7.723e-01   0.065 0.948035
## faixa_salarioD    -6.919e-01  1.277e+00  -0.542 0.587986
## AtrasoSim          5.607e-01  3.282e-01   1.709 0.087504 .
## ValorEmprestimo   -1.186e-02  3.704e-03  -3.203 0.001362 **
## QtdaPagas          1.058e+00  3.874e-01   2.731 0.006313 **
## salario           -9.071e-06  1.916e-05  -0.474 0.635829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 556.26  on 472  degrees of freedom
## Residual deviance: 516.81  on 461  degrees of freedom
## AIC: 540.81
##
## Number of Fisher Scoring iterations: 4
```

- Retirando a variável faixa_salario pois já tenho a variável de salário

```
logistica <- glm(default1 ~ Sexo +
                           EstadoCivil +
                           NumerodeFilhos +
                           TempodeResidencia +
                           Atraso +
                           ValorEmprestimo +
                           QtdaPagas+
                           salario, family = binomial, data = ds_analise)
summary(logistica)
```

```
##
## Call:
## glm(formula = default1 ~ Sexo + EstadoCivil + NumerodeFilhos +
##     TempodeResidencia + Atraso + ValorEmprestimo + QtdaPagas +
##     salario, family = binomial, data = ds_analise)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3714  -1.0880   0.6518   0.8204   1.4254
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.853e+00  6.382e-01   4.471 7.79e-06 ***
## SexoMasculino     -3.315e-01  2.446e-01  -1.355 0.175465
## EstadoCivil        8.797e-04  1.072e-01   0.008 0.993450
## NumerodeFilhos     1.203e-02  5.952e-02   0.202 0.839786
## TempodeResidencia  8.044e-03  1.109e-02   0.725 0.468148
## AtrasoSim          5.681e-01  3.252e-01   1.747 0.080694 .
## ValorEmprestimo   -1.220e-02  3.690e-03  -3.306 0.000946 ***
## QtdaPagas          1.093e+00  3.860e-01   2.831 0.004643 **
## salario           -1.369e-05  6.589e-06  -2.077 0.037809 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 556.26  on 472  degrees of freedom
## Residual deviance: 523.18  on 464  degrees of freedom
## AIC: 541.18
##
## Number of Fisher Scoring iterations: 4
```

- Retirando TempodeResidencia

```
logistica <- glm(default1 ~ Sexo +
                           EstadoCivil +
                           NumerodeFilhos +
                           Atraso +
                           ValorEmprestimo +
                           QtdaPagas+
                           salario, family = binomial, data = ds_analise)
summary(logistica)
```

```
##
## Call:
## glm(formula = default1 ~ Sexo + EstadoCivil + NumerodeFilhos +
##     Atraso + ValorEmprestimo + QtdaPagas + salario, family = binomial,
##     data = ds_analise)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2789  -1.0952   0.6626   0.8152   1.4533
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.944e+00  6.266e-01   4.698 2.62e-06 ***
## SexoMasculino   -3.194e-01  2.440e-01  -1.309 0.190634
## EstadoCivil     -2.905e-03  1.070e-01  -0.027 0.978342
## NumerodeFilhos   9.428e-03  5.934e-02   0.159 0.873751
## AtrasoSim        5.732e-01  3.249e-01   1.764 0.077726 .
## ValorEmprestimo -1.228e-02  3.678e-03  -3.337 0.000846 ***
## QtdaPagas        1.101e+00  3.845e-01   2.864 0.004178 **
## salario         -1.369e-05  6.591e-06  -2.077 0.037792 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 556.26  on 472  degrees of freedom
## Residual deviance: 523.72  on 465  degrees of freedom
## AIC: 539.72
##
## Number of Fisher Scoring iterations: 4
```

```
# Obtive um ganho no modelo e também não a vejo como significante no contexto de negócio
```

- Retirando Atraso

```
logistica <- glm(default1 ~ Sexo +
                           EstadoCivil +
                           NumerodeFilhos +
                           ValorEmprestimo +
                           QtdaPagas+
                           salario, family = binomial, data = ds_analise)
summary(logistica)
```

```
##
## Call:
## glm(formula = default1 ~ Sexo + EstadoCivil + NumerodeFilhos +
##     ValorEmprestimo + QtdaPagas + salario, family = binomial,
##     data = ds_analise)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2841  -1.1088   0.6604   0.8189   1.4622
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.220e+00  6.113e-01   5.268 1.38e-07 ***
## SexoMasculino   -2.873e-01  2.429e-01  -1.183 0.236922
## EstadoCivil     -1.260e-02  1.065e-01  -0.118 0.905854
## NumerodeFilhos   3.570e-03  5.917e-02   0.060 0.951896
## ValorEmprestimo -1.221e-02  3.670e-03  -3.327 0.000879 ***
## QtdaPagas        1.075e+00  3.834e-01   2.805 0.005027 **
## salario         -1.478e-05  6.549e-06  -2.258 0.023963 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 556.26  on 472  degrees of freedom
## Residual deviance: 527.07  on 466  degrees of freedom
## AIC: 541.07
##
## Number of Fisher Scoring iterations: 4
```

```
# Obtive um ganho no modelo
```

- Vou realizar a acuracia do modelo

```
prob <- predict(logistica, ds_analise, type = 'response')
resultado <- if_else(prob >= 0.50, 1,0)
target <- ds_analise$default1
tabela <- table(target, resultado)
acuracia <- ((tabela[1] + tabela[4]) / sum(tabela))
acuracia
```

```
## [1] 0.744186
```

```
# Acuracia de 74,41%
```

b. Criar um modelo de árvore de decisão com a variável target default e interpretar as regras do modelo e verificar o quanto o modelo teve de acurácia.

```
library(rpart)
library(rattle)
```

```
## Warning: package 'rattle' was built under R version 4.1.2
```

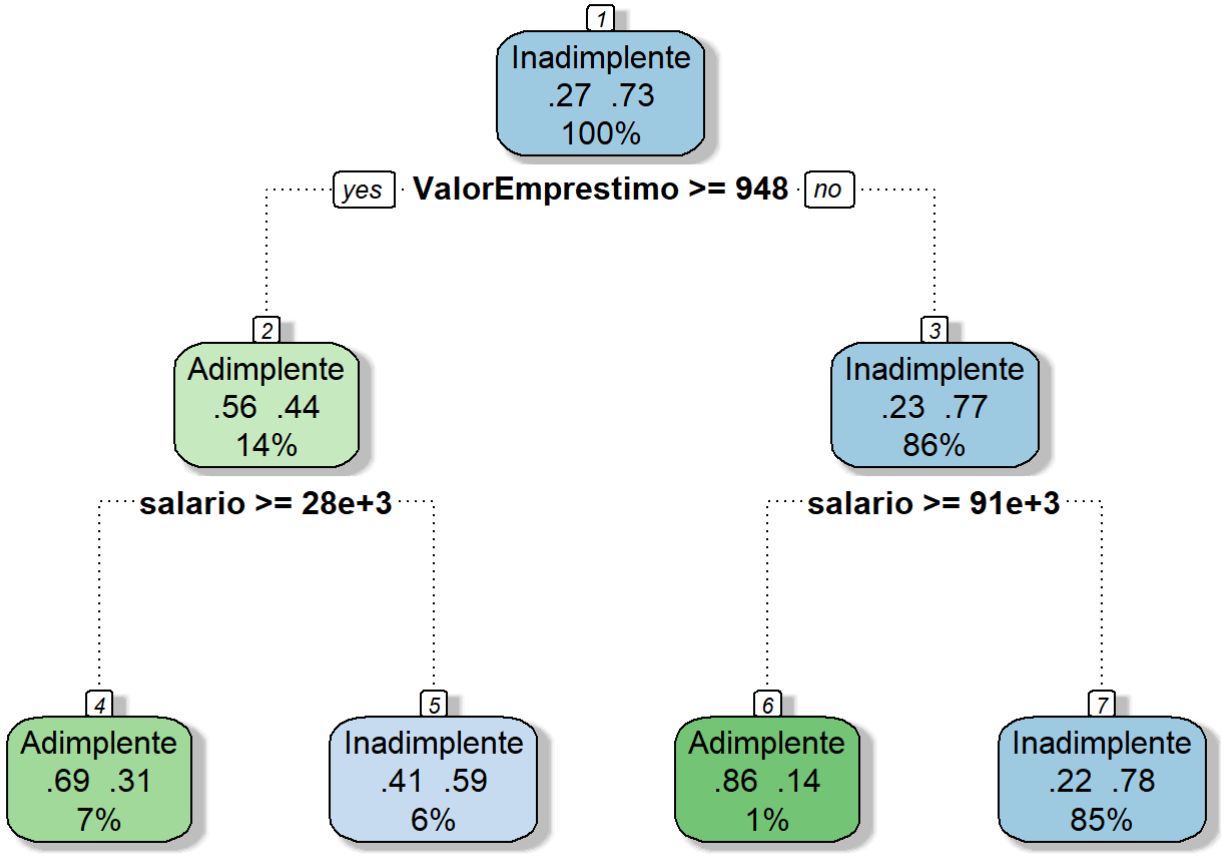
```
## Carregando pacotes exigidos: bitops
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
mytree = rpart(default ~ Sexo +
                EstadoCivil +
                NumerodeFilhos +
                ValorEmprestimo +
                QtdaPagas+
                salario,
                data = ds_analise,
                method="class")
mytree
```

```
## n= 473
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 473 130 Inadimplente (0.2748414 0.7251586)
##   2) ValorEmprestimo>=947.5 64  28 Adimplente (0.5625000 0.4375000)
##     4) salario>=27825 35  11 Adimplente (0.6857143 0.3142857) *
##     5) salario< 27825 29  12 Inadimplente (0.4137931 0.5862069) *
##   3) ValorEmprestimo< 947.5 409  94 Inadimplente (0.2298289 0.7701711)
##     6) salario>=90937.5 7   1 Adimplente (0.8571429 0.1428571) *
##     7) salario< 90937.5 402  88 Inadimplente (0.2189055 0.7810945) *
```

```
fancyRpartPlot(mytree)
```



Verificando acuracia da arvore de decisão

```
ds_analise$probArvore = predict(mytree, newdata = ds_analise, type="prob")
ds_analise$resultadoArvore = predict(mytree, newdata = ds_analise, type="class")
tabela_arvore  = table(ds_analise$default, ds_analise$resultadoArvore)
tabela_arvore
```

```
##
##                Adimplente Inadimplente
##   Adimplente           30          100
##   Inadimplente         12          331
```

```
acuracia_arvore <- (tabela_arvore[1] + tabela_arvore[4])/sum(tabela_arvore)
acuracia_arvore
```

```
## [1] 0.7632135
```

```
# Acuracia de 76,32%
```

c. Verifique qual modelo teve o melhor desempenho e justifique sua resposta.

```
#A arvore de decisão teve um melhor desempenho 76,32% de acuracia, já a regressão logistica #
teve 74,41%.
#A arvore apresentou um ganho de 1,91% de acuracia em relação à regressão logistica
```

d. Explique o porquê você escolheu cada uma das variáveis

```
# Sexo -> Para a analise poderia influenciar, se for homem ou mulher. Pensando no contexto ne
gócio
# Estadocivil -> Uma pessoa casada por exemplo poderia levar a uma inadimplência do que uma p
essoa  solteira. Pensando no contexto negócio
# NumerodeFilhos -> Se uma pessoa possui mais filhos a renda acaba sendo comprometida.Pensand
o no contexto negócio
# TempodeResidencia -> Resolvi tirar pois não vejo que influenciaria. Pensando no contexto ne
gócio
# Atraso -> Retirei pois não mostrou muito significativa
# TempodeServiço -> Exclui esta variável pois o modelo regressão logistica não estava converg
indo, e                    estava muito correlacionada com o default
# comprometido_de_renda -> Exclui esta variável pois o modelo regressão logistica não estava
 convergindo
# faixa_salario -> Exclui pois optei por ficar com a variável salário e tive um ganho modelo
# Conta -> Exclui pois não estava apresentando nenhum dado (1 not defined because of singular
ities)
```