Analise Czech

Análise Inadimplência

Um banco quer melhorar os seus servicos.

Como exemplo, os gerentes tem apenas uma idéia vaga, de quem é um bom cliente e de quem é um mal cliente.

Os gerentes não possuem nenhuma pergunta especifica, então a análise ocorrerá com enfoque na exploração, tentando responder uma questão hipotética.

- O que, dentro das característoas de um empréstimo, e do seu perfil, influencia a inadimplencia de um cliente?

Análise de inadimplência do cliente

Um bom cliente, iremos definir que é o cliente que realiza empréstimos e paga os mesmos. Cliente adimplente. Um mal cliente, iremos definir que é o cliente que realizar emprestimos e não paga os mesmos. Cliente inadimplente.

As situações de um empréstimo podem ser:

- 'A' Contrato de emprestimo finalizado sem problema.
- 'B' Contrato de emprestimo finalizado, porém o emprestimo não foi pago.
- 'C' Contrato de emprestimo em andamento, com pagamentos em dia.
- 'D' Contrato de emprestimo em andamento, com pagamentos em atraso.

A pergunta que se quer responder é: Conseguimos determinar, dado um cliente e as características de um empréstimo, a probabilidade dele se tornar inadimplente? Este valor pode ser importante para o cálculo de possíveis perdas, valor do spread, etc.

Poderíamos também determinar quem é um bom cliente por outros prismas, mas para esta análise iremos focar na relação de cliente e emprestimos.

Leitura de dados

O banco possue dados históricos, de transações, emprestimos, geolocalizacao, uso do cartão e outros. Os dados foram limpos (data wrangling) e encontram-se em uma base MySQL.

Para mais informações sobre o modelo de dados, vide documento **PKDD'99 Discovery Challenge Guide to the Financial Data Set**.

Estes dados serão utilizados para a modelagem do problema e tentativa de construção de um modelo preditivo logístico.

```
## Loading required package: DBI
```

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

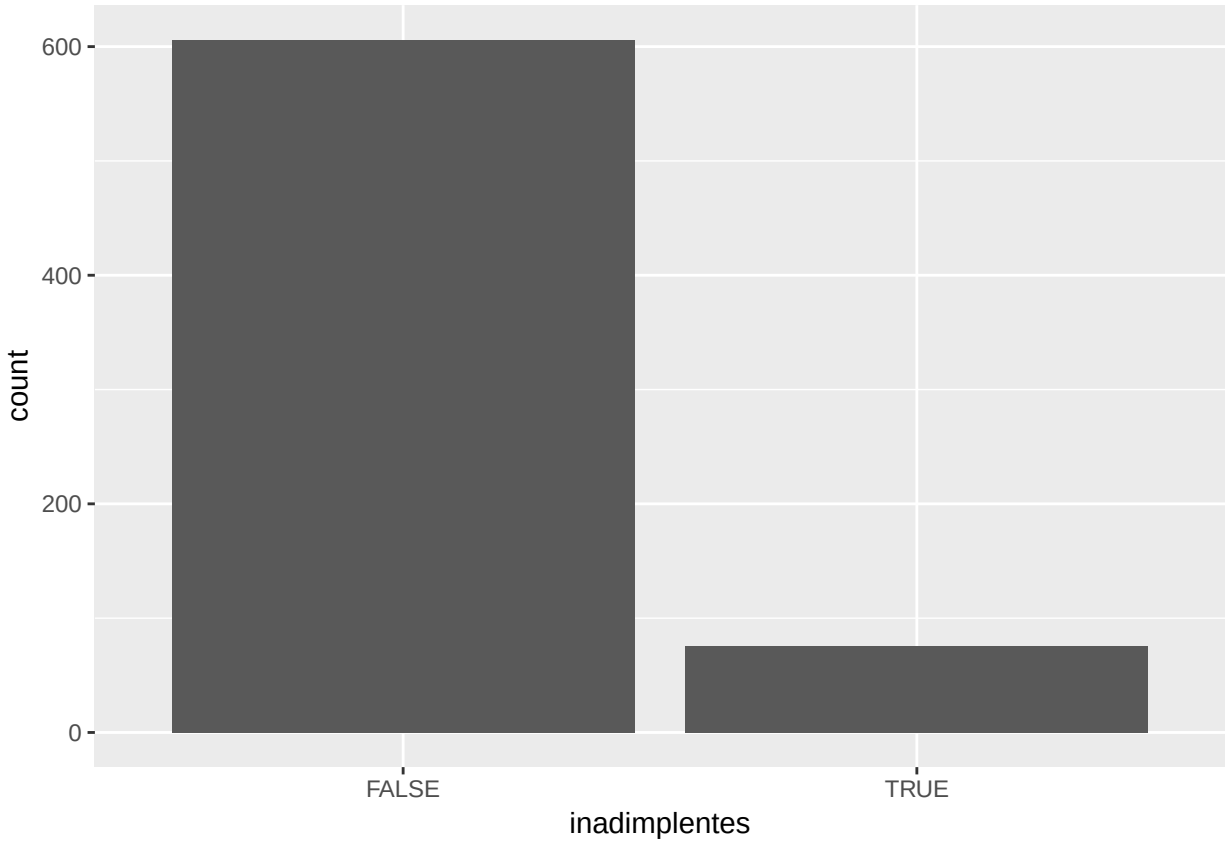
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

Análise Exploratória

Os empréstimos possuem a distribuição a seguir:

```
loan <- dbReadTable(con, "loan")
loan <- loan %>% mutate(inadimplentes = (status == "B" | status == "D"))
loan %>% ggplot(aes(x=inadimplentes)) + geom_bar()
```



Procurando correlações

Montante do empréstimo influencia a inadimplência?

- H0 : O montante do empréstimo não influencia a inadimplência.
- HA : O montante do empréstimo influencia a inadimplência.

```
model_amount <- glm(inadimplentes ~ amount, data = loan)
summary(model_amount)
```

```
##
## Call:
## glm(formula = inadimplentes ~ amount, data = loan)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.31590  -0.12891  -0.08565  -0.06161   0.94531
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.098e-02  1.986e-02   2.064   0.0394 *
## amount      4.653e-07  1.050e-07   4.431 1.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09652288)
##
##     Null deviance: 67.531  on 681  degrees of freedom
## Residual deviance: 65.636  on 680  degrees of freedom
## AIC: 344.93
##
## Number of Fisher Scoring iterations: 2
```

```
melhor_modelo <- model_amount
```

Logo, podemos ver que o valor do emprestimo influencia a probabilidade de inadimplencia (p-value = 0,000355) - existe uma forte corelação entre o amount e a adimplencia.

A duração dos emprestimos

- H0 : A duração dos empréstimos não influência a inadimplencia
- HA : A duração dos empréstimos influência a inadimplencia

```
model_duration <- glm(inadimplentes ~ durantion, data = loan)
summary(model_duration)
```

```
##
## Call:
## glm(formula = inadimplentes ~ durantion, data = loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.12263  -0.11692  -0.11120  -0.09977   0.90023
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0940581  0.0284807   3.303  0.00101 **
## durantion   0.0004762  0.0007070   0.674  0.50079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09924377)
##
##     Null deviance: 67.531  on 681  degrees of freedom
## Residual deviance: 67.486  on 680  degrees of freedom
## AIC: 363.89
##
## Number of Fisher Scoring iterations: 2
```

Podemos visualizar que a duração não influencia (p-value = 0.146)

E ambos, montante e duração, juntos, melhoram o modelo preditivo?

```
novo_modelo <- glm(inadimplentes ~ durantion + amount, data = loan)
summary(novo_modelo)
```

```
##
## Call:
## glm(formula = inadimplentes ~ durantion + amount, data = loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.35446  -0.12959  -0.09310  -0.05778   0.97400
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.202e-02  2.797e-02   3.289  0.00106 **
## durantion   -2.265e-03  8.783e-04  -2.579  0.01011 *
## amount       6.743e-07  1.323e-07   5.097 4.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09572713)
##
##     Null deviance: 67.531  on 681  degrees of freedom
## Residual deviance: 64.999  on 679  degrees of freedom
## AIC: 340.28
##
## Number of Fisher Scoring iterations: 2
```

```
melhor_modelo <- novo_modelo
```

O valor e a duração juntos influenciam a inadimplencia!

O valor das parcelas a serem pagas influenciam?

- H0 : O valor das parcelas não influência a inadimplencia
- HA : O valor das parcelas influência a inadimplencia

```
novo_modelo <- glm(inadimplentes ~ payments, data = loan)
summary(novo_modelo)
```

```
##
## Call:
## glm(formula = inadimplentes ~ payments, data = loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.25973  -0.14159  -0.09139  -0.04902   0.95389
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.783e-03  2.540e-02   0.110    0.913
```

```
## payments    2.593e-05  5.358e-06   4.839 1.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09600451)
##
##     Null deviance: 67.531  on 681  degrees of freedom
## Residual deviance: 65.283  on 680  degrees of freedom
## AIC: 341.26
##
## Number of Fisher Scoring iterations: 2
```

O valor da parcela influencia a inadimplencia!

Mas devemos incluir a valor no modelo candidato ?

```
novo_modelo <- glm(inadimplentes ~ amount + durantion + payments, data = loan)
summary(novo_modelo)
```

```
##
## Call:
## glm(formula = inadimplentes ~ amount + durantion + payments,
##     data = loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.33948  -0.13750  -0.08974  -0.05306   0.96721
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.798e-02  6.087e-02   0.953    0.341
## amount       4.980e-07  3.096e-07   1.609    0.108
## durantion   -1.508e-03  1.490e-03  -1.012    0.312
## payments     7.891e-06  1.253e-05   0.630    0.529
##
## (Dispersion parameter for gaussian family taken to be 0.09581229)
##
##     Null deviance: 67.531  on 681  degrees of freedom
## Residual deviance: 64.961  on 678  degrees of freedom
## AIC: 341.88
##
## Number of Fisher Scoring iterations: 2
```

Neste modelo, todos os p-values são maiores que 0,05. Como individualmente eles possuem relevancia, vamos reduzir a dois preditores.

```
step_model <- step(novo_modelo)
```

```
## Start:  AIC=341.88
## inadimplentes ~ amount + durantion + payments
##
##             Df Deviance    AIC
## - payments   1   64.999 340.28
## - durantion  1   65.059 340.91
## <none>           64.961 341.88
## - amount     1   65.209 342.48
##
```

```
## Step:  AIC=340.28
## inadimplentes ~ amount + durantion
##
##             Df Deviance    AIC
## <none>           64.999 340.28
## - durantion  1   65.636 344.93
## - amount     1   67.486 363.89
```

```
summary(step_model)
```

```
##
## Call:
## glm(formula = inadimplentes ~ amount + durantion, data = loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.35446  -0.12959  -0.09310  -0.05778   0.97400
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.202e-02  2.797e-02   3.289  0.00106 **
## amount       6.743e-07  1.323e-07   5.097 4.48e-07 ***
## durantion   -2.265e-03  8.783e-04  -2.579  0.01011 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09572713)
##
##     Null deviance: 67.531  on 681  degrees of freedom
## Residual deviance: 64.999  on 679  degrees of freedom
## AIC: 340.28
##
## Number of Fisher Scoring iterations: 2
```

Reduzindo-se assim para apenas duas variáveis, Montante do emprestimo e duração do emprestimo, temos o melhor modelo até o momento.

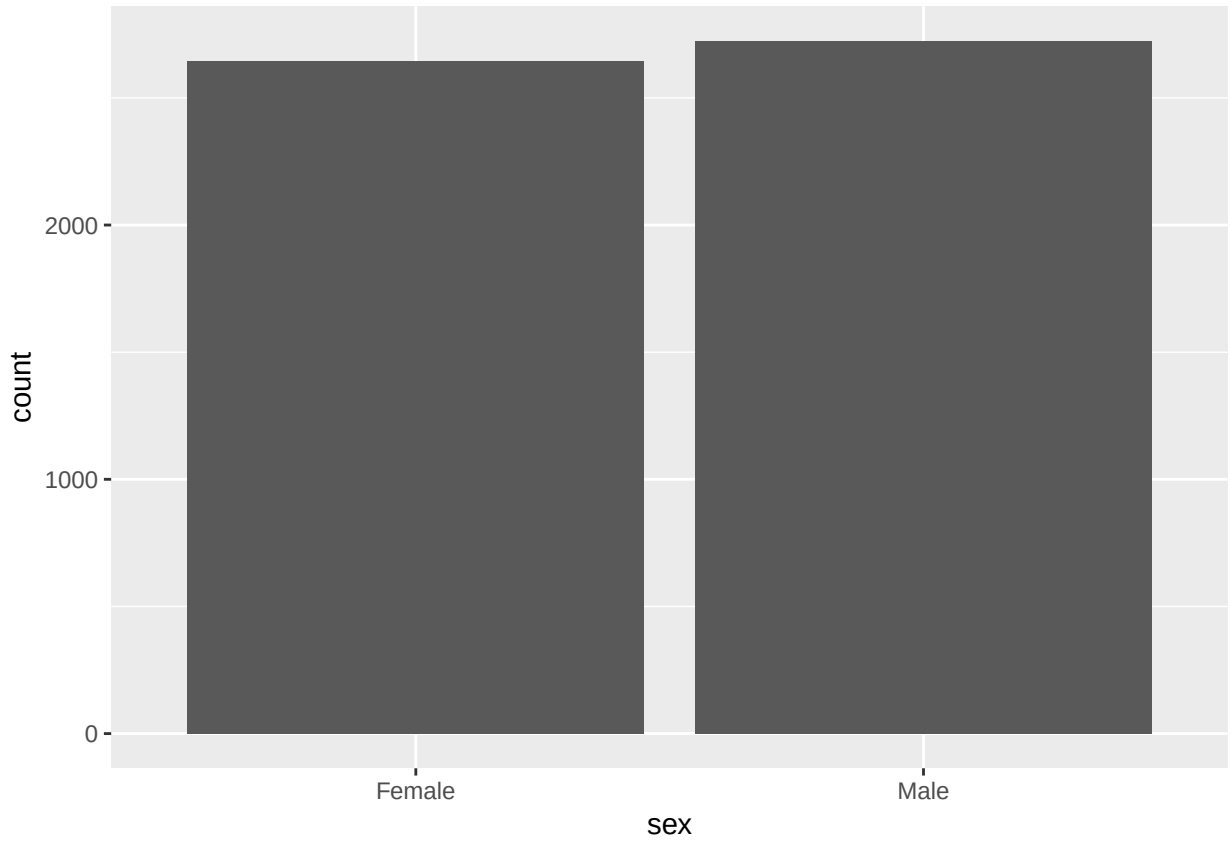
```
melhor_modelo <- step_model
```

O Sexo do cliente influencia na inadimplência?

- H0 : O sexo do cliente não influencia a inadimplência
- HA : O sexo do cliente influencia a inadimplência

O sexo entre os clientes se distribui como

```
account_client %>% ggplot(aes(x=sex)) + geom_bar()
```



Logo, temos uma distribuição quase igual entre ambos os sexos, mas em relacão a inadimplencia, como se comportam ambos os sexos?

```
account_loan <- account_client %>% merge(loan, by = "account_id")
novo_modelo <- glm(inadimplentes ~ sex , data = account_loan)
summary(novo_modelo)
```

```
##
## Call:
## glm(formula = inadimplentes ~ sex, data = account_loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.09832  -0.09832  -0.08537  -0.08537   0.91463
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09832    0.01416   6.943 7.75e-12 ***
## sexMale     -0.01296    0.02011  -0.644     0.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.08361336)
##
##     Null deviance: 69.016  on 826  degrees of freedom
## Residual deviance: 68.981  on 825  degrees of freedom
## AIC: 298.68
##
```

```
## Number of Fisher Scoring iterations: 2
```

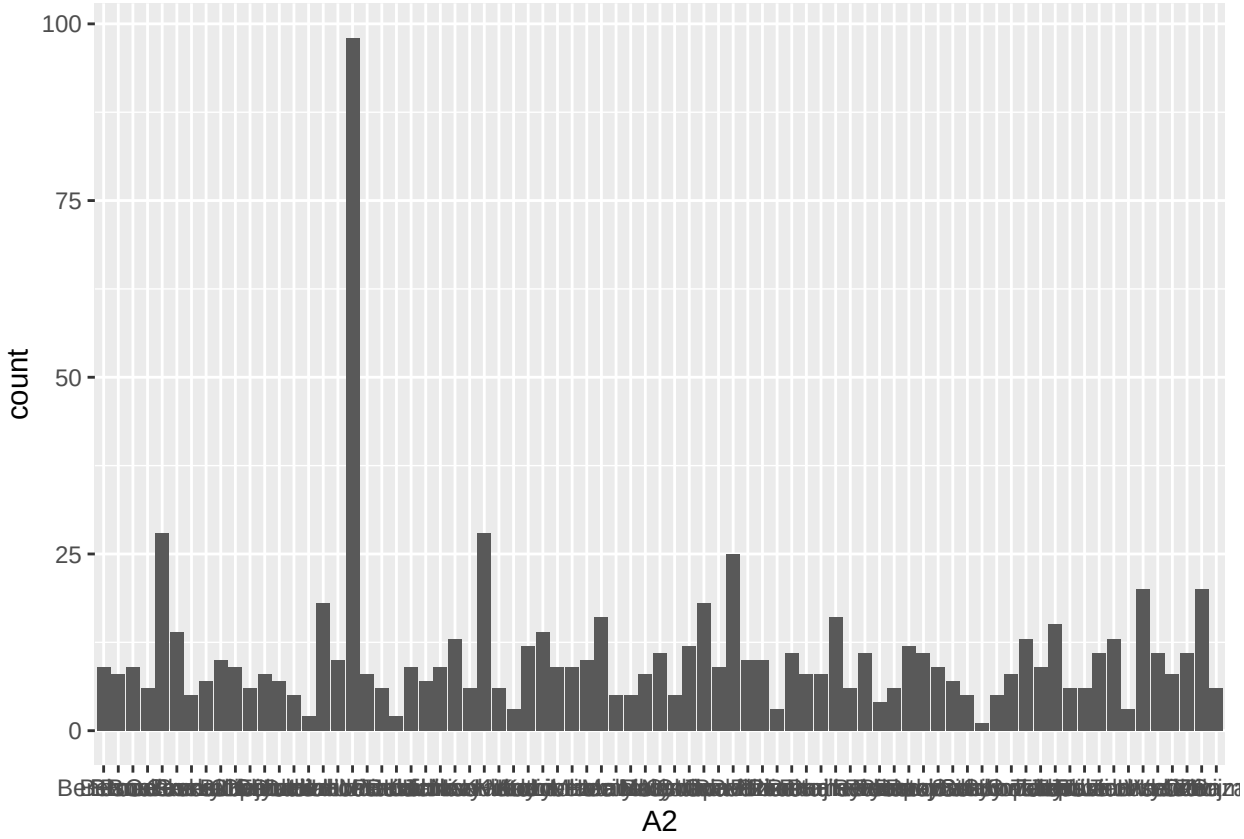
Podemos verificar que o sexo do cliente não possui relação (alto p-value) com a inadimplencia.

Região geográfica influencia?

- H0 : As regiões não influenciam na inadimplencia
- H1 : As regiões inflenciam na inadimplencia

Vejamos o total dos emprestimos por região

```
geo_loan <- merge(account_loan, demograph, by.x = c("district_id.y"), by.y = c("district_id"))
geo_loan %>% ggplot(aes(x=A2)) + geom_bar()
```



Analisemos a quantidade de emprestimos por regiao

```
 x <- geo_loan %>% group_by(A2) %>%
  summarise(total = n())
```

E a quantidade de inadimplentes por região

```
  geo_loan %>% filter(inadimplentes) %>%group_by(A2) %>%
  summarise(total = n())
```

```
## # A tibble: 43 x 2
##    A2               total
##    <chr>            <int>
##  1 Beroun               2
##  2 Blansko              1
##  3 Breclav              1
```

```
##  4 Brno - mesto         5
##  5 Brno - venkov        1
##  6 Bruntal              3
##  7 Ceske Budejovice     1
##  8 Cesky Krumlov        1
##  9 Chrudim              2
## 10 Domazlice            1
## # ... with 33 more rows
```

Existe alguma relação entre região e inadimplencia, i.é, alguma região onde a inadimplencia é melhor?

```
novo_modelo <- glm(inadimplentes ~ A2, data =  geo_loan)
summary(novo_modelo)
```

```
##
## Call:
## glm(formula = inadimplentes ~ A2, data = geo_loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.60000  -0.11111  -0.06122   0.00000   0.95000
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.039e-14  9.410e-02   0.000 1.000000
## A2Beroun                2.500e-01  1.372e-01   1.823 0.068775 .
## A2Blansko               1.111e-01  1.331e-01   0.835 0.404019
## A2Breclav               1.667e-01  1.488e-01   1.120 0.262993
## A2Brno - mesto          1.786e-01  1.082e-01   1.651 0.099193 .
## A2Brno - venkov         7.143e-02  1.206e-01   0.592 0.553882
## A2Bruntal               6.000e-01  1.575e-01   3.811 0.000150 ***
## A2Ceska Lipa           -2.092e-14  1.423e-01   0.000 1.000000
## A2Ceske Budejovice      1.000e-01  1.297e-01   0.771 0.440971
## A2Cesky Krumlov         1.111e-01  1.331e-01   0.835 0.404019
## A2Cheb                 -1.209e-14  1.488e-01   0.000 1.000000
## A2Chomutov             -1.961e-14  1.372e-01   0.000 1.000000
## A2Chrudim               2.857e-01  1.423e-01   2.008 0.044968 *
## A2Decin                -2.147e-14  1.575e-01   0.000 1.000000
## A2Domazlice             5.000e-01  2.207e-01   2.266 0.023755 *
## A2Frydek - Mistek       5.556e-02  1.152e-01   0.482 0.629910
## A2Havlickuv Brod        1.000e-01  1.297e-01   0.771 0.440971
## A2Hl.m. Praha           6.122e-02  9.833e-02   0.623 0.533691
## A2Hodonin              -2.068e-14  1.372e-01   0.000 1.000000
## A2Hradec Kralove        1.667e-01  1.488e-01   1.120 0.262993
## A2Jablonec n. Nisou    -2.226e-14  2.207e-01   0.000 1.000000
## A2Jesenik               2.222e-01  1.331e-01   1.670 0.095362 .
## A2Jicin                -1.892e-14  1.423e-01   0.000 1.000000
## A2Jihlava              -2.078e-14  1.331e-01   0.000 1.000000
## A2Jindrichuv Hradec     7.692e-02  1.224e-01   0.628 0.529941
## A2Karlovy Vary          3.333e-01  1.488e-01   2.240 0.025359 *
## A2Karvina               1.071e-01  1.082e-01   0.990 0.322252
## A2Kladno                3.333e-01  1.488e-01   2.240 0.025359 *
## A2Klatovy              -2.000e-14  1.882e-01   0.000 1.000000
## A2Kolin                 8.333e-02  1.245e-01   0.669 0.503422
## A2Kromeriz              1.429e-01  1.206e-01   1.184 0.236614
```

```
## A2Kutna Hora           2.222e-01  1.331e-01   1.670 0.095362 .
## A2Liberec             -2.115e-14  1.331e-01   0.000 1.000000
## A2Litomerice          -2.189e-14  1.297e-01   0.000 1.000000
## A2Louny               -2.032e-14  1.176e-01   0.000 1.000000
## A2Melnik              -1.924e-14  1.575e-01   0.000 1.000000
## A2Mlada Boleslav       2.000e-01  1.575e-01   1.270 0.204418
## A2Most                 1.250e-01  1.372e-01   0.911 0.362452
## A2Nachod              -2.053e-14  1.269e-01   0.000 1.000000
## A2Novy Jicin          -2.068e-14  1.575e-01   0.000 1.000000
## A2Nymburk             -1.986e-14  1.245e-01   0.000 1.000000
## A2Olomouc              1.667e-01  1.152e-01   1.446 0.148552
## A2Opava                3.333e-01  1.331e-01   2.505 0.012463 *
## A2Ostrava - mesto      1.200e-01  1.097e-01   1.094 0.274521
## A2Pardubice           -2.081e-14  1.297e-01   0.000 1.000000
## A2Pelhrimov           -2.180e-14  1.297e-01   0.000 1.000000
## A2Pisek               -2.035e-14  1.882e-01   0.000 1.000000
## A2Plzen - jih         -2.000e-14  1.269e-01   0.000 1.000000
## A2Plzen - mesto        1.250e-01  1.372e-01   0.911 0.362452
## A2Plzen - sever       -2.089e-14  1.372e-01   0.000 1.000000
## A2Prachatice           6.250e-02  1.176e-01   0.531 0.595333
## A2Praha - vychod      -2.024e-14  1.488e-01   0.000 1.000000
## A2Praha - zapad        9.091e-02  1.269e-01   0.716 0.473922
## A2Prerov              -2.128e-14  1.696e-01   0.000 1.000000
## A2Pribram             -1.969e-14  1.488e-01   0.000 1.000000
## A2Prostejov            1.667e-01  1.245e-01   1.339 0.181016
## A2Rakovnik             9.091e-02  1.269e-01   0.716 0.473922
## A2Rokycany             1.111e-01  1.331e-01   0.835 0.404019
## A2Rychnov nad Kneznou -2.135e-14  1.423e-01   0.000 1.000000
## A2Semily              -2.179e-14  1.575e-01   0.000 1.000000
## A2Sokolov              1.000e+00  2.976e-01   3.361 0.000817 ***
## A2Strakonice           6.000e-01  1.575e-01   3.811 0.000150 ***
## A2Sumperk             -2.146e-14  1.372e-01   0.000 1.000000
## A2Svitavy              2.308e-01  1.224e-01   1.885 0.059794 .
## A2Tabor                2.222e-01  1.331e-01   1.670 0.095362 .
## A2Tachov               1.333e-01  1.190e-01   1.120 0.262993
## A2Teplice             -2.112e-14  1.488e-01   0.000 1.000000
## A2Trebic               1.667e-01  1.488e-01   1.120 0.262993
## A2Trutnov             -2.094e-14  1.269e-01   0.000 1.000000
## A2Uherske Hradiste    -2.100e-14  1.224e-01   0.000 1.000000
## A2Usti nad Labem      -2.112e-14  1.882e-01   0.000 1.000000
## A2Usti nad Orlici      5.000e-02  1.133e-01   0.441 0.659150
## A2Vsetin               9.091e-02  1.269e-01   0.716 0.473922
## A2Vyskov              -2.110e-14  1.372e-01   0.000 1.000000
## A2Zdar nad Sazavou    -2.115e-14  1.269e-01   0.000 1.000000
## A2Zlin                 5.000e-02  1.133e-01   0.441 0.659150
## A2Znojmo               1.667e-01  1.488e-01   1.120 0.262993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.07969304)
##
##     Null deviance: 69.016  on 826  degrees of freedom
## Residual deviance: 59.770  on 750  degrees of freedom
## AIC: 330.14
```

```
##
## Number of Fisher Scoring iterations: 2
```

Devido a quantidade de regiões, podemos procurar as que mais influenciam

```
step(novo_modelo)
```

```
## Start:  AIC=330.14
## inadimplentes ~ A2
##
##        Df Deviance    AIC
## - A2   76   69.016 297.09
## <none>      59.770 330.14
##
## Step:  AIC=297.09
## inadimplentes ~ 1
```

```
##
## Call:  glm(formula = inadimplentes ~ 1, data = geo_loan)
##
## Coefficients:
## (Intercept)
##      0.0919
##
## Degrees of Freedom: 826 Total (i.e. Null);  826 Residual
## Null Deviance:       69.02
## Residual Deviance: 69.02     AIC: 297.1
```

Assim, a região também não influencia a inadimplencia

Salário influencia?

- H0 : O Salário médio da região do não cliente influencia na inadimplencia
- HA : O Salário médio da região do cliente influencia na inadimplencia

```
novo_modelo <- glm(inadimplentes ~ A11, data =  geo_loan)
summary(novo_modelo)
```

```
##
## Call:
## glm(formula = inadimplentes ~ A11, data = geo_loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.09974  -0.09722  -0.09432  -0.08620   0.92595
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.468e-01  7.314e-02   2.007   0.0451 *
## A11         -5.798e-06  7.655e-06  -0.757   0.4491
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.0835973)
##
##     Null deviance: 69.016  on 826  degrees of freedom
```

```
## Residual deviance: 68.968  on 825  degrees of freedom
## AIC: 298.52
##
## Number of Fisher Scoring iterations: 2
```

Também não há influencia estatistica relevante em relação ao salário médio da região do cliente.

Conclusão

Os melhores preditores para se determinar a inadimplencia ou não, são a duração e o montante do emprestimo.

```
summary(melhor_modelo)
```

```
##
## Call:
## glm(formula = inadimplentes ~ amount + durantion, data = loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.35446  -0.12959  -0.09310  -0.05778   0.97400
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.202e-02  2.797e-02   3.289  0.00106 **
## amount       6.743e-07  1.323e-07   5.097 4.48e-07 ***
## durantion   -2.265e-03  8.783e-04  -2.579  0.01011 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09572713)
##
##     Null deviance: 67.531  on 681  degrees of freedom
## Residual deviance: 64.999  on 679  degrees of freedom
## AIC: 340.28
##
## Number of Fisher Scoring iterations: 2
```

Porém ainda pode-se explorar outras variáveis relativas ao cliente, como volume de transações, a quanto tempo ele é cliente, etc.

```
## [[1]]
## [1] TRUE
```