Analise Czech

Objetivo

Um banco quer melhor os seus servicos.

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

Os gerentes não possuem nenhum pergunta especifica, então a análise ocorrera com enfoque na exploração, tentando responder algumas hipotéticas questões

- O que, dentro das característoas de um emprestimo, influencia a inadimplencia de um cliente
- A quem deve-se ofertar/reduzir o cartão de acordo com o com região / crédito / perfil do cliente.
- ofertar/reduzir crédito de acordo com região / crédito / perfil do cliente
- marketing para abrir contas em determinada região.

Analise de inadimplencia do cliente

Um bom cliente, iremos definir que é o cliente que realizar emprestimos 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 emprestimo 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: Dado um cliente, qual a probabilidade de quando o emprestimo for finalizado, ele se tornar inadimplente ou ter contrato default?

Podemos 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 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 construção de um modelo preditivo

```
## 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

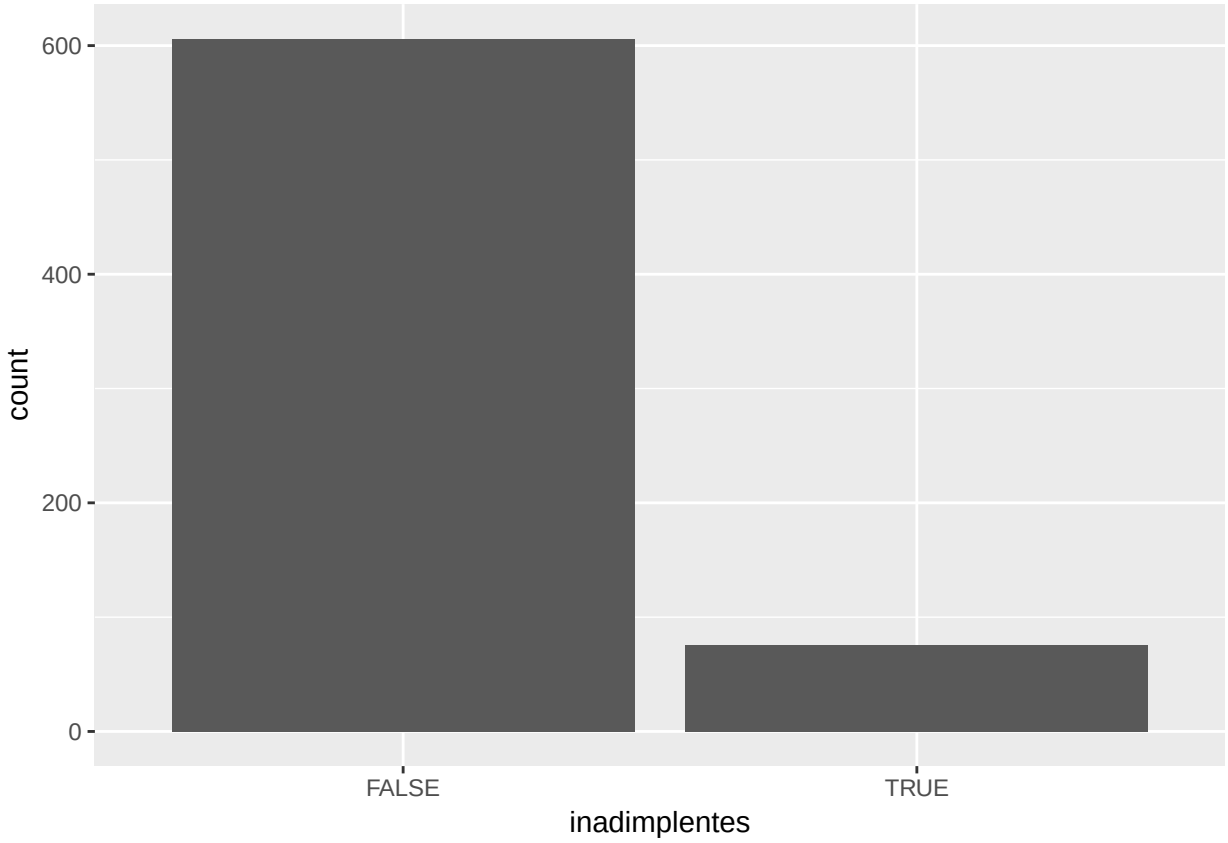
Os empréstimos, possuem a distribuição a seguir:

```
loan <- dbReadTable(con, "loan")
```

```
## Warning in .local(conn, statement, ...): Decimal MySQL column 3 imported as
## numeric
```

```
## Warning in .local(conn, statement, ...): Decimal MySQL column 6 imported as
## numeric
```

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



Correlações, apenas utilizando variáveis de empréstimo.

Montante do empréstimo influência?

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 duraçãoo 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 juntos?

```
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 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)
```

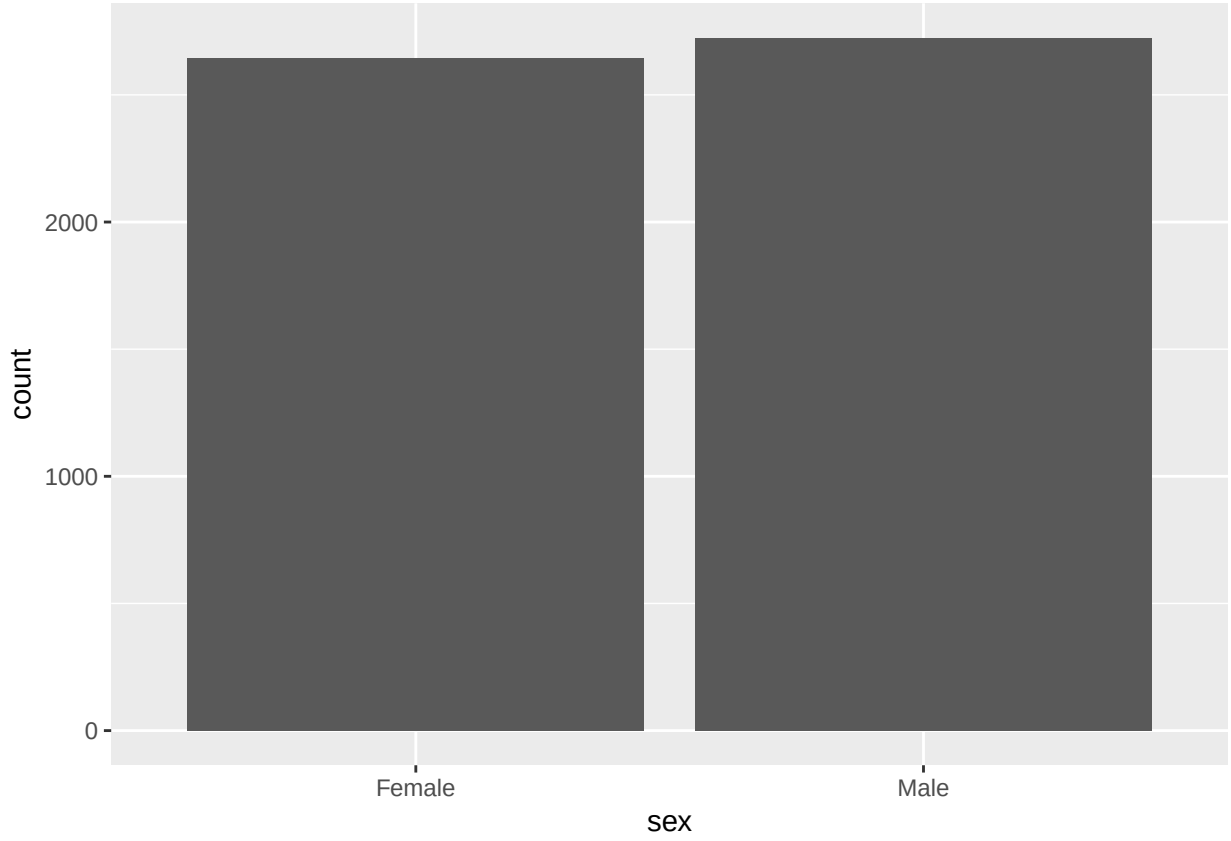
```
##
## Call:
## glm(formula = inadimplentes ~ amount + durantion + payments,
##     data = loan)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.33948  -0.13750  -0.08974  -0.05306   0.96721
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (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, assim, optamos por não utilizar o valor da prestacao.

O Sexo do cliente influencia na inadimplencia?

O sexo entre os clientes se distribui como

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



Logo, temos uma distribuição quase equilibrado 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 com a inadimplencia ou não do mesmo.

A idade do cliente influencia na inadimplencia?