MBA Executivo em Business Analytics e Big Data

Modelagem Estatística Avançada

Estudo de Caso: Concessão de Crédito

dataset: german credit data

Professor Gustavo Corrêa Mirapalheta

Alunos:
Antonio Henrique Trotta
João Ignácio de Almeida Junior
Rafael Nonato

Abril de 2021

Introdução/ Resumo

Trabalho final da matéria de Modelagem Estatística Avançada ministrada pelo professor Gustavo Corrêa Mirapalheta. Nesse estudo de caso, o objetivo é auxiliar um gerente de empréstimo bancário a desenvolver um modelo que forneça a probabilidade de um cliente em potencial ficar inadimplente (*default* de crédito), com maior precisão, e definir uma probabilidade de corte (aquela a partir da qual o cliente terá o crédito negado) que irá maximizar o lucro esperado para a carteira de clientes do banco.

Para apurar o lucro, conforme as orientações do material, levamos em conta as regras de negócios descritas na tabela abaixo. O custo de oportunidade (proveniente do lucro de um crédito bom e do prejuízo decorrente de um crédito ruim) será conforme definido abaixo:

	Previsão	
Real	Bom	Ruim
Bom	$ 100	$ -100
Ruim	$ -500	$ -

Para os clientes previstos como “bons pagadores”: Se o cliente for de fato um bom pagador, o resultado será + $100; por outro lado, se na realidade for um mau pagador, o resultado para o banco será de - $500.

Para os clientes previstos como “maus pagadores”: Para esses clientes o empréstimo será negado, porém se o cliente for na realidade um bom pagador, o resultado para o banco será de - $100 por não emprestar o dinheiro a alguém que de fato pagaria corretamente; por outro lado, em se comprovando de fato um mau pagador, o resultado será $0, pois, nesse cenário, não houve prejuízo ao banco.

Base de dados: conforme orientação, utilizamos a base de dados *german_data,* disponível em: *https://raw.githubusercontent.com/gustavomirapalheta/classes_datasets/master/german_data.csv*

Contexto:

a) **Definições de negócio para o caso**

A partir da base de dados definida acima, realizaremos a regressão linear logística a fim de comparar a probabilidade de êxito dos modelos que serão descritos a seguir.

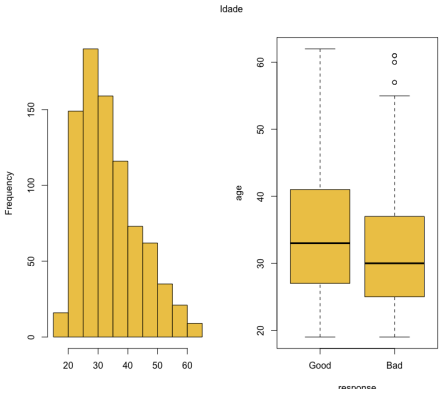
b) **Produto a ser ofertado**

Nosso objetivo é identificar e ofertar um modelo seguro de previsão que identifique os bons pagadores dos maus pagadores. Dessa forma, maximizando o lucro e diminuindo o risco de prejuízos com os empréstimos realizados.

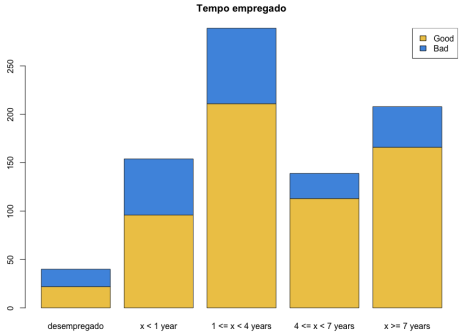
c) Análise de variáveis

A seguir analisamos as principais variáveis utilizadas em nossa avaliação. Removemos os *outliers* para fazermos nossas análises:

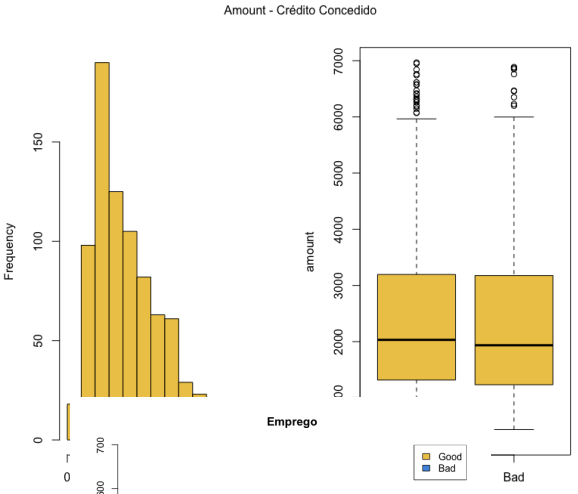
1) **Idade:** Conforme podemos observar no gráfico ao lado, existe uma concentração entre os tomadores de empréstimo na faixa dos 30 anos, e a partir de então a população de tomadores de empréstimo cai gradativamente. Não identificamos desvios relevantes entre bons e maus pagadores. Variável quantitativa.



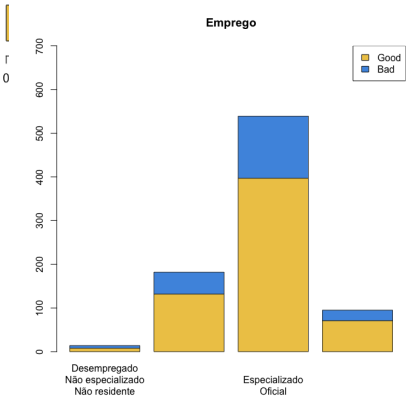
2) **Tempo empregado:** Aproximadamente 35% da população está empregada entre 1 e 4 anos; chamamos atenção para 5% da população que está desempregada; possivelmente maus pagadores. Variável qualitativa.



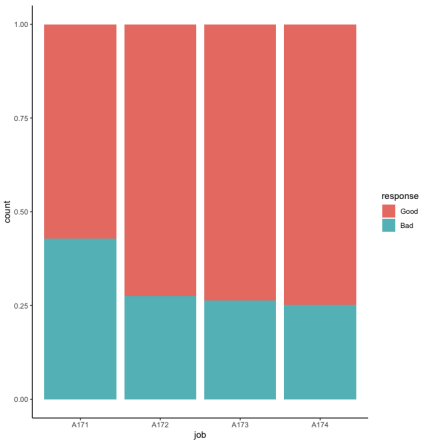
3) **Valor do crédito concedido:** Conforme podemos observar no gráfico ao lado, a grande maioria dos empréstimos não supera o valor de $5.000. Os valores médios são substancialmente os mesmos, tanto para bons quanto maus pagadores. Variável numérica.



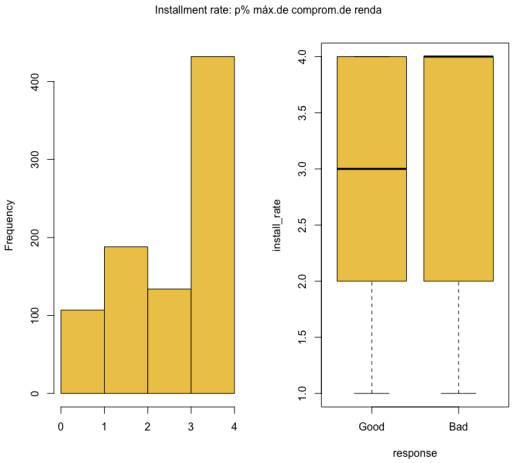
4) **Tipo de emprego:** Mais de 60% dos tomadores de empréstimos possuem emprego de especialista técnico. Variável qualitativa.



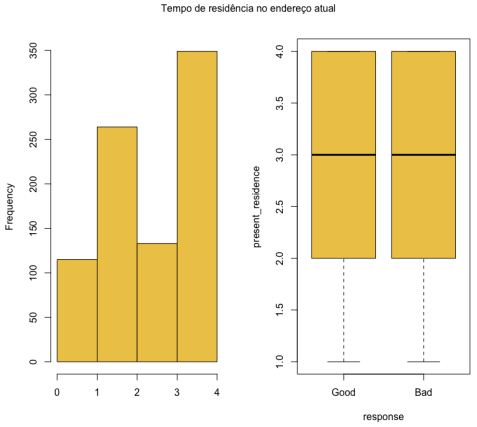
**Removido - Proporções muito próximas entre os itens observados em nossa amostra; pode aumentar a correlação.*



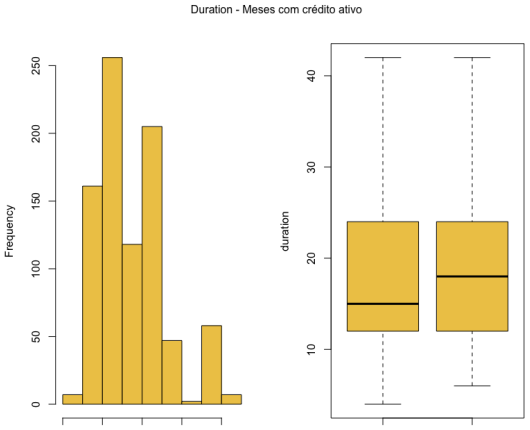
5) **Installment Rate:** Indica o quanto o empréstimo representa do salário de cada tomador de empréstimo



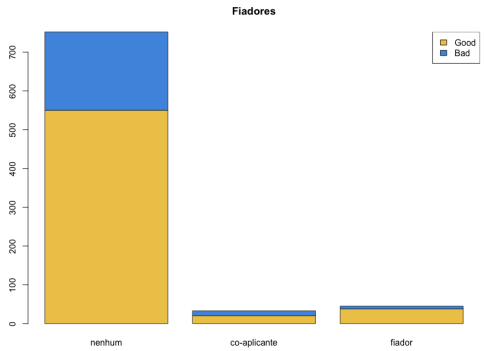
6) T**empo de residência atual:** Essa informação demonstra o quanto tempo o tomador de empréstimo reside no endereço atual. **Removido - Não relevante para a variável resposta.*



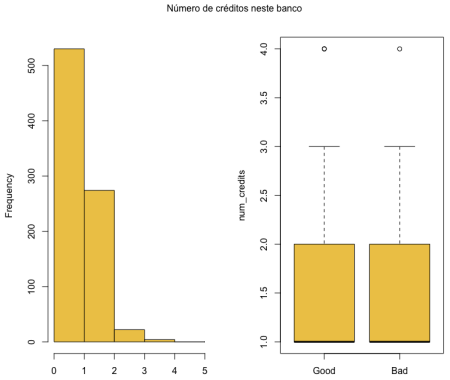
7) **Duração do empréstimo:** Ao lado, podemos verificar que a maioria dos empréstimos não ultrapassa os 25 meses; porém podemos observar que para os maus pagadores, a média é em torno de 35 meses, no entanto, para os bons pagadores, a média fica em torno dos 18 meses.



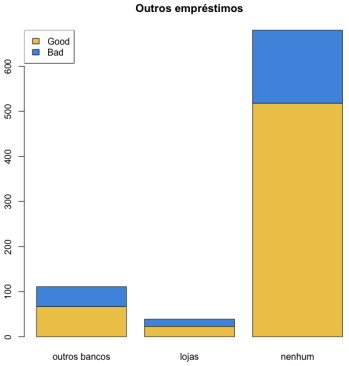
8) **Fiador:** Informação qualitativa, no entanto, quase a totalidade dos tomadores de empréstimo não utiliza um fiador no processo. Essa informação, portanto, não nos apóia de forma eficaz no modelo para detecção de maus pagadores.
*Removido V*ariável tem baixa distribuição, podendo aumentar correlacao*



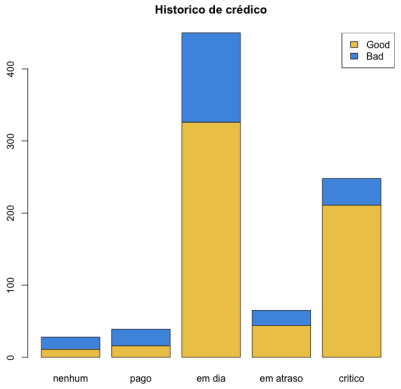
9) **Número de empréstimos:** Mais de 60% da amostra possui apenas 1 empréstimo nesse banco, e mais de 90% possui até 2 empréstimos. Não observamos distinções para bons ou maus pagadores. *Removido *Não é relevante para a variável resposta.*



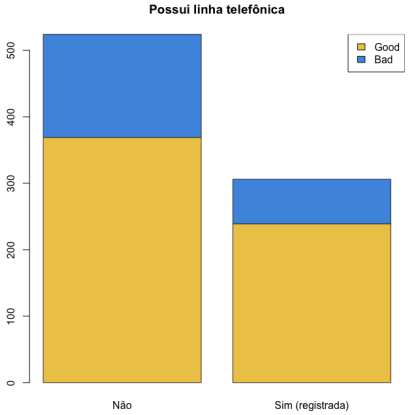
10) **Other Installment:** Outros empréstimos mantidos pelos tomadores.
**Removido - variável possui baixa distribuição, podendo aumentar correlação*



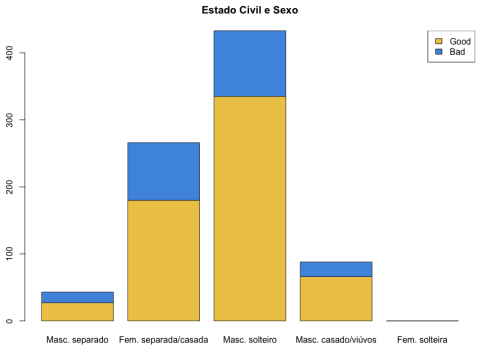
11) **Histórico de pagamentos:** Conforme podemos observar, aproximadamente 60% dos tomadores têm quitado suas dívidas no prazo; os títulos classificados como “em atraso” e “críticos” indicam os maus pagadores em nssa população.



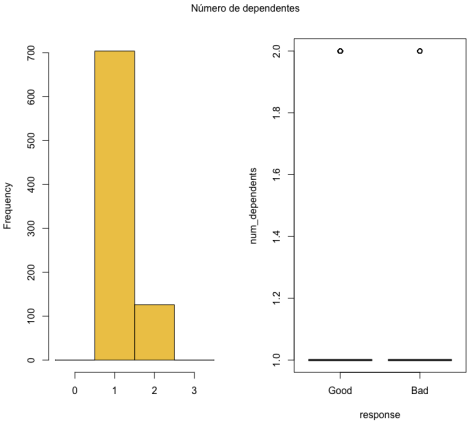
12) Linha telefônica: Essa variável indica se o tomador de empréstimo possui linha telefônica ou não. Como podemos observar no gráfico, a proporção de maus pagadores é maior na população que não possui linha telefônica.



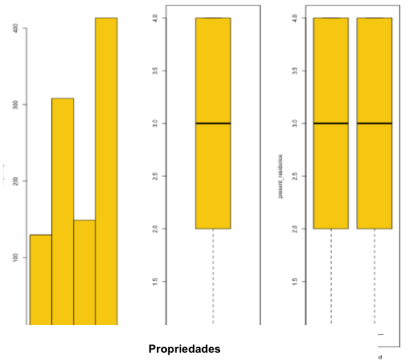
12) **Estado civil e sexo:** Aproximadamente 55% de nossa população é composta de homens solteiros;30% são mulheres separadas ou casadas. Os 15% restantes são homens casados, viúvos e separados. Praticamente não há mulheres solteiras tomadoras de empréstimos.



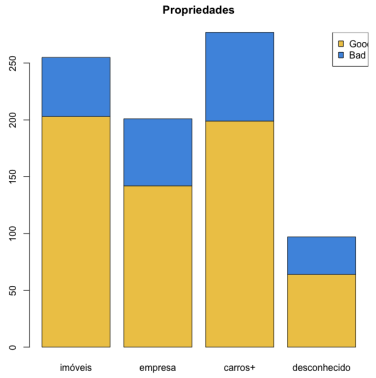
13) **Número de dependentes:** Mais de 80% dos tomadores de empréstimo possui um dependente; a parcela restante possui 2 dependentes. *Removido -Nào é relevante para variável resposta.



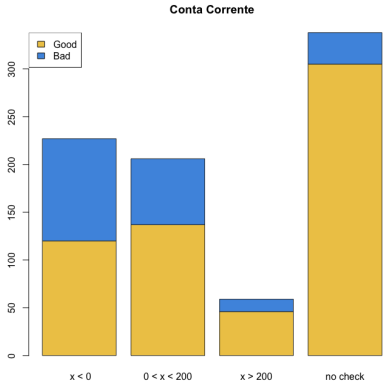
14) **Anos em que se habita a moradia atual:** mais de 40% dos tomadores de empréstimo habitam suas casas há 4 anos. Aproximadamente 30%, por 2 anos. A média é de 3 anos e não observamos distinção entre bons e maus pagadores.



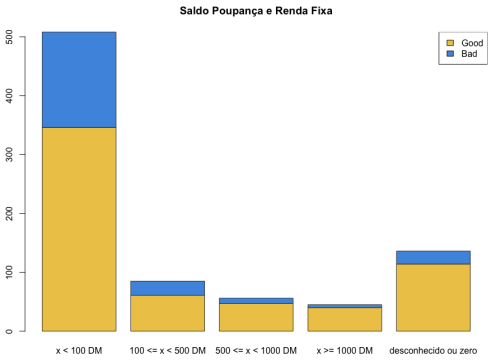
15) **Propriedades:** Como forma de garantia aos empréstimos a pagar, os tomadores possuem propriedades da seguinte forma: aproximadamente 35% possuem carros; aproximadamente 30% possuem imóveis; 23% possuem empresas e os demais não informaram.



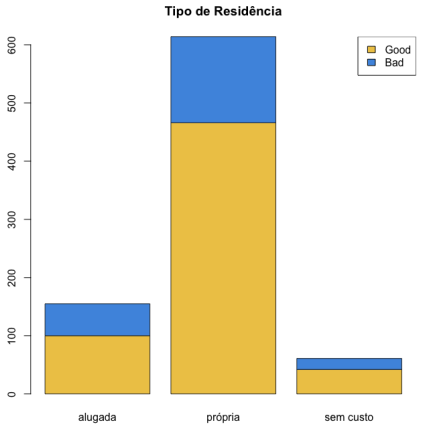
16) **Saldo disponível em conta:** Essa informação está diretamente relacionada à capacidade de os tomadores de empréstimos quitarem suas dívidas. Quase 30% estão com saldo negativo, o que levanta dúvida quanto ao pagamento dos empréstimos. 40% também não informou o saldo.



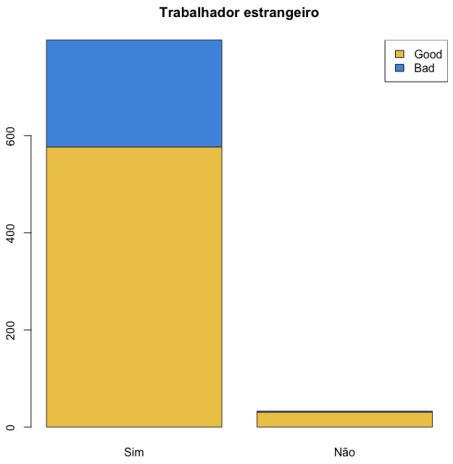
17) **Saldo em aplicações:** Mais uma informação relevante para avaliar a qualidade dos pagadores e a capacidade de quitar suas dívidas. 60% da população possui menos de $ 100 DM aplicados; no entanto, aproximadamente 20% possui entre $100 e $1.000 DM, aumenta a probabilidade de quitarem suas dívidas no prazo. Outros 20% têm $0 ou não informaram.



18) **Tipo de residência:** 70% da população mora em residência própria; aproximadamente 20% mora de aluguel e outros 10% moram sem custo (família).
*Removido - *variável tem baixa distribuição, podendo aumentar correlação*

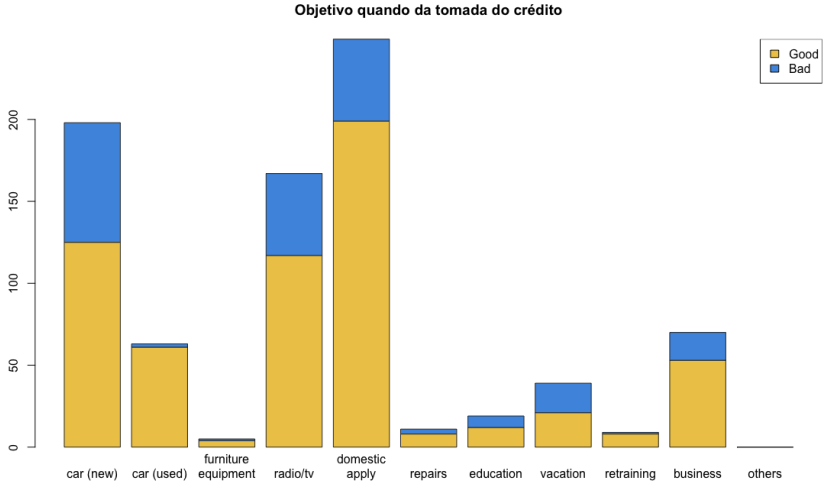


19) **Trabalhador Estrangeiro**
**Removido - baixa distribuição dos dados, podendo levar a alta correlação*



20) **Objetivo do empréstimo:** Como podemos observar no gráfico abaixo, a maioria da população utiliza o empréstimo obtido para fins domésticos (30%), para aquisição de carro novo (25%), rádio e tv (18%) e carro usado (10%). Outras utilidades, menos representativas, são: aquisição de móveis/ equipamentos, férias, estudos e negócios.

Realizamos correlação dessa informação com o histórico de pagamentos e identificamos abaixo os bons e maus pagadores de empréstimos por natureza.



d) Modelos desenvolvidos

Durante o exercício de análise das principais variáveis descritas acima, desenvolvemos nosso modelo com o objetivo de prever os bons e maus pagadores com maior chance de êxito. A princípio consideramos em nosso modelo todas as variáveis propostas (modelo denominado “FULL”). Nesse primeiro exercício, não removemos os outliers, obtivemos um AIC (Akaike Information Criterion) de 725.19.
No mesmo modelo, após remoção dos outliers o modelo, atingimos o AIC 543.18.
Em uma terceira abordagem, removemos as variáveis por intuição, chegou-se a base pré-processada com apenas 12 variáveis, onde ao rodar este modelo, chegamos ao AIC: 542.44.
Numa quarta tentativa, com ajuda do algoritmo de remoção variáveis correlacionadas no método STEPWISE, chegamos ao modelo com AIC: 513.44

Para justificar a escolha do modelo verificamos o conjunto de variáveis a seguir:

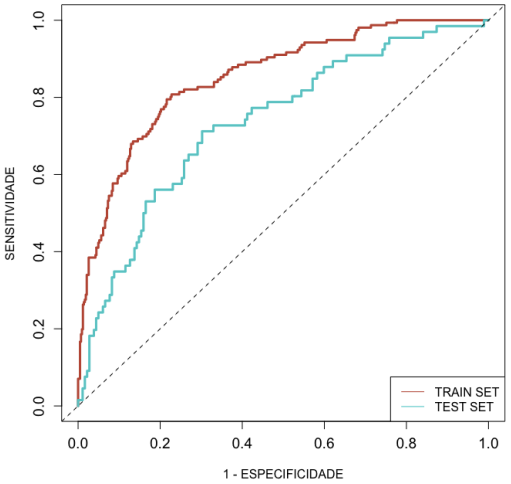
```
> # MODELO 2 - VARIAVEIS INTUITIVAS
> Y_PROB_TRAIN <- predict(modelo_2, type = 'response')
> Y_PROB_TEST  <- predict(modelo_2, newdata = testing_set_2, type = 'response')
> head(Y_PROB_TRAIN)
         5          9         10         13         14         16
0.80715440 0.01206569 0.91498920 0.25812302 0.57297066 0.41855814
> # MODELO INTUITIVO STEP
> Y_PROB_TRAIN_STEP <- predict(modelo_STEP, type = 'response')
> Y_PROB_TEST_STEP  <- predict(modelo_STEP, newdata = testing_set, type = 'response')
> head(Y_PROB_TRAIN_STEP)
         5          9         10         13         14         16
0.82638261 0.01580629 0.88567901 0.22454751 0.49072893 0.37435518
```

Com base no modelo de STEPWISE, com apenas 5 variáveis, explica aproximadamente 82,5% do teste.
Por esta série de avaliações para chegar no modelo final, que segue descrito no tópico a seguir.

e) Modelo escolhido

Definimos o modelo *Stepwise*.

MODELO FINAL = MODELO STEPWISE



CURVA ROC DO MODELO

f) Resultados estatísticos e conclusão

```
Confusion Matrix and Statistics

          Reference
Prediction Good Bad
      Good  180   2
      Bad    63   3

               Accuracy : 0.7379
                 95% CI : (0.6785, 0.7915)
    No Information Rate : 0.9798
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0489

 Mcnemar's Test P-Value : 9.911e-14

            Sensitivity : 0.74074
            Specificity : 0.60000
         Pos Pred Value : 0.98901
         Neg Pred Value : 0.04545
             Prevalence : 0.97984
         Detection Rate : 0.72581
   Detection Prevalence : 0.73387
      Balanced Accuracy : 0.67037

       'Positive' Class : Good
```

```
> # VALOR TOTAL PREVISTO PELO MODELO
> (confusao$table[1] * 100) + (confusao$table[3] * (-500) ) + (confusao$table[4] * 0 ) + (confusao$table[2] * (-100) )
[1] 10700
```

Em nosso código é possível verificar a realização de 6 modelos distintos através dos quais fomos refinando nosso modelo gradualmente, seja pela avaliação das variáveis de forma individual, seja pela remoção dos outliers. Concluímos por fim pelo modelo denominado *Stepwise*, com AIC de 513.44.

MODELO STEP WISE
Null deviance: 676.63 on 581 degrees of freedom
Residual deviance: 461.44 on 556 degrees of freedom
AIC: 513.44

Com o ponto de corte em 71%, atingimos uma acurácia de 66%. Mas devido à perda potencial dos falsos positivos, precisamos ajustar o corte, pois o valor previsto neste ponto de corte seria um lucro de aproximadamente $5.900.

Ponto de corte final: em 89%, a acurácia fica em 67% e o lucro é de $10.700.

O ponto de corte máximo seria de 97%, com lucro de $11.000, porém acurácia de 36%, portanto, sendo de baixa confiança.

Códigos R

O script do R utilizado para as conclusões acima descritas está acessível pelo link abaixo:

https://colab.research.google.com/drive/1PWdqTJNaHARxVD_iHyzz8CnifnVdCfoE?usp=sharing