Aplicação de regressão logística para a classificação de clientes inadimplentes

Universidade Federal da Paraíba - UFPB

Abril 2024

Sumário

Introdução

A inadimplência em dívidas de cartão de crédito é um problema financeiro significativo que afeta tanto os consumidores quanto as instituições financeiras. A capacidade de prever e mitigar a inadimplência é crucial para manter a estabilidade financeira e reduzir os riscos associados ao crédito.

A importância desse problema reside na sua ampla repercussão econômica e social. A inadimplência pode levar a uma série de consequências adversas, tanto para os consumidores quanto para as instituições financeiras. Para os consumidores, a inadimplência pode resultar em penalidades financeiras, danos à sua pontuação de crédito e dificuldades financeiras prolongadas. Para as instituições financeiras, a inadimplência representa perdas financeiras, aumento dos custos de empréstimos e uma diminuição da confiança do mercado.

Além disso, a inadimplência pode ser um indicador de problemas financeiros mais amplos, como desemprego, recessão econômica ou má administração financeira pessoal. Portanto, entender os fatores que contribuem para a inadimplência em dívidas de cartão de crédito é essencial para desenvolver estratégias eficazes de mitigação de riscos e políticas de crédito responsáveis.

Dessa forma, com base num banco de dados com informações de crédito de clientes, o presente trabalho busca utilizar técnicas de regressão logística para classificar clientes quanto a sua situação financeira. A base de dados contém informações sobre o saldo do cliente, a renda e se o cliente é ou não um estudante.

Metodologia

A base de dados

O banco de dados de dados utilizado contém informações de inadimplência de clientes de cartão de crédito. Na base de dados temos 4 variáveis, das quais 2 são nominais e 2 são numéricas. Dessa forma, temos as seguintes variáveis:

- **Inadimplência** : Esta variável representa se o cliente entrou em inadimplência no cartão de crédito ou não. É uma das variáveis nominais com dois níveis: "Não", indicando que o cliente não entrou em inadimplência, e "Sim", indicando que o cliente entrou em inadimplência. Esta será a variável dependente para a modelagem com regressão logística, em que o objetivo será classificar se um cliente entrará em inadimplência em sua dívida no cartão de crédito com base em outras características no conjunto de dados.
- **Estudante**: A variável *Estudante* é a segunda variável nominal e ela indica se o cliente é ou não estudante. Ela possui dois níveis: "Não", indicando que o cliente não é estudante, e "Sim", indicando que é estudante. Esta variável independente pode ser importante para discriminar o comportamento de um cliente que é estudante para um que não é estudante, pois isso poderia significar diferentes perfis de risco.
- **Saldo**: O Saldo é uma das variáveis contínua numérica representa o saldo médio que o cliente tem remanescente em seu cartão de crédito após efetuar o pagamento mensal. O saldo reflete o valor da dívida no cartão de crédito que o cliente carrega em média. Saldos mais altos podem indicar níveis mais elevados de dívida e potencialmente maior risco de inadimplência.
- **Renda**: Por último, temos a renda do cliente, que é a última variável numérica do banco de dados. A renda é um fator importante na avaliação de solvência, já que indivíduos com rendas mais altas podem ser mais propensos a pagar dívidas pontualmente. Maiores rendas podem indicar um menor risco de inadimplência, embora isso possa variar dependendo de outros fatores.

Além disso, este banco de dados contém 10.000 observações. 20% foram reservados para teste, enquanto os outros 80% foram usados para ajustar um modelo de regressão logística.

Recursos computacionais

Para realizar a modelagem, a exploração dos dados e todas as outras análises que estão presente nesse trabalho, foi utilizada a linguagem de programação R. Como produto da linguagem R, foram utilizados pacotes para modelagem estatística e criação de gráficos. Para a modelagem estatística foi utilizado o framework *tidymodels* e para a visualização de gráficos, foi utilizado o *ggplot2*. Além disso, foi também utilizado o Quarto, que serve para fazer apresentações e documentos de escrita, o que é o caso desse documento.

Análise exploratória

Um dos primeiros passos para qualquer estudo estatístico, é fazer a análise exploratória dos dados. Dessa forma, esse foi o primeiro passo tomado nesse projeto, utilizando medidas de tendência central e de dispersão, como média, mediana, desvio-padrão e coeficiente de correlação, além da utilização de gráficos e tabelas com o objetivo de caracterizar e compreender melhor os eventos em questão. Além disso, foram elaborados gráficos para observar o comportamento da variável resposta e das variáveis preditoras.

Regressão logística

Construção do modelo e métricas para sua validação

Uma das etapas mais importantes para a validação de um modelo, é a escolha de métricas para analisar a sua performance e formas de verificar se as variáveis que estão sendo utilizadas para a modelagem são estatisticamente significantes. Nesse caso, para a análise da regressão logística, foram escolhidas métricas que são geradas a partir de uma matriz de confusão e a escolha das variáveis foi feita pelo teste de significância dos coeficientes da regressão. Por fim, foi analisado como o modelo de regressão estava performando durante a classificação.

Sensitividade

A sensitividade, também chamado de frequência de verdadeiros positivos, é calculado como a quantidade verdadeiros positivos dividido pela quantidade total de positivos. Dessa forma, temos que:

$$Sensitividade = \frac{TP}{TP + FN}$$

em que TP é a quantidade de verdadeiros positivos e FN é a quantidade de falsos negativos. Ainda, a melhor sensitividade seria 1, enquanto a pior seria 0.

Especificidade

Frequência de verdadeiros negativos ou especificidade, é calculado como a quantidade de verdadeiros negativos dividido pelo total de negativos. Portanto, temos que:

$$Especificidade = \frac{TN}{TN + FP}$$

o TN é a quantidade de verdadeiros negativos e FP é a quantidade de falsos positivos. Temos também que a melhor sensitividade para o modelo de regressão logística se apróxima de 1, enquanto a pior se aproxima de 0.

Acurácia

A acurácia, que também foi utilizada para analisar a performance do modelo, é definida como a quantidade de todas as corretas classificações dividido pela total do banco de dados. Assim, temos que:

$$Acurcia = \frac{TP + TN}{TP + TN + FN + FP}$$

É importante saber também que quanto mais próximo de 1 está a acurácia, mais o modelo está classificando bem.

Área Sob a Curva ROC (AUC)

AUC significa Área Sob a Curva ROC. É uma métrica bastante conhecida para avaliar o desempenho de um modelo de classificação binária. A curva ROC plota a frequência de verdadeiros positivos **(sensibilidade)** contra a taxa de falsos positivos **(1 - especificidade)** para diferentes valores de limiar. A métrica de AUC quantifica o poder discriminativo geral do modelo em todos os valores de limiar possíveis. A interpretação do AUC é feita de forma bastante simples. Um AUC igual a 1 indica um classificador perfeito que separa perfeitamente as instâncias positivas e negativas. A curva ROC alcança o canto superior esquerdo, significando alta sensibilidade e especificidade.

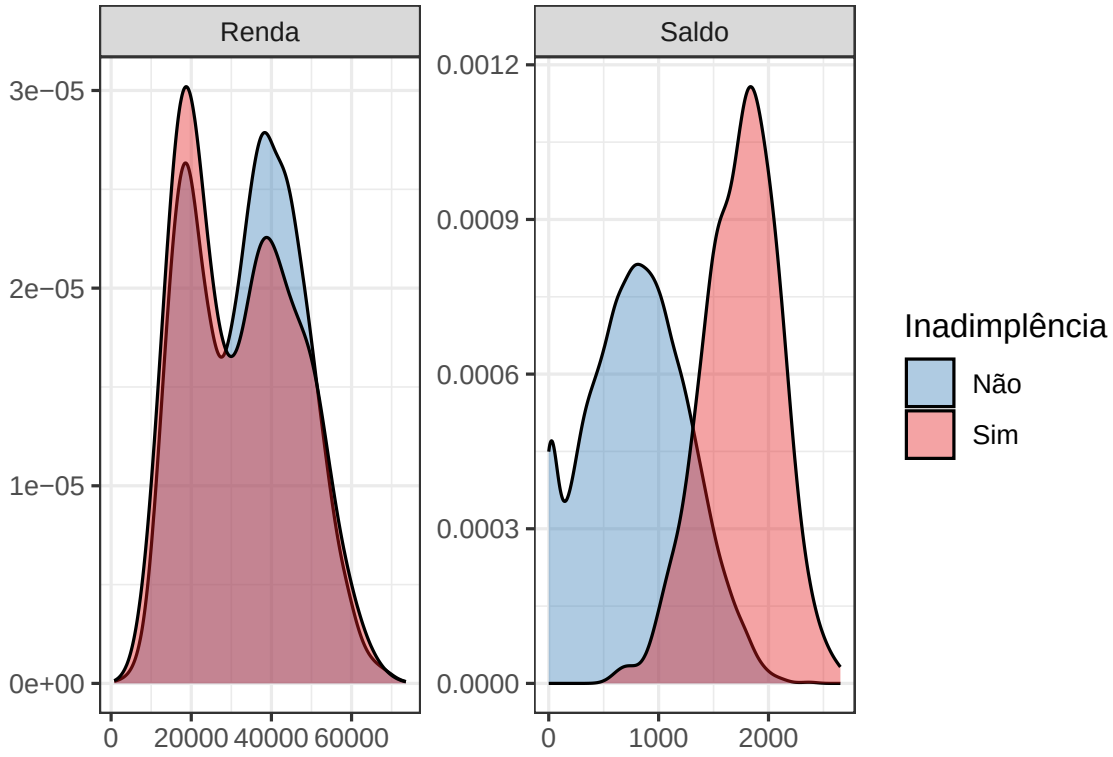
Análise dos diagnóstico

A análise de diagnósticos é um dos passos mais importantes para confirmar a situação do modelo, suas suposições e o quão bem ele está classificando ou fazendo estimativas. Uma das suposições a ser verificada pela análise de dignósticos é a normalidade dos resíduos, que pode ser feita utilizando um gráfico *QQ-plot* ou testes como *Shapiro-Wilk* ou *Lilliefors*. Ainda, com a análise dos diagnósticos, é possível verificar a suposição de heterocedasticidade, identificar pontos aberrantes, alavanca ou influentes.

Resultados

Exploração dos dados

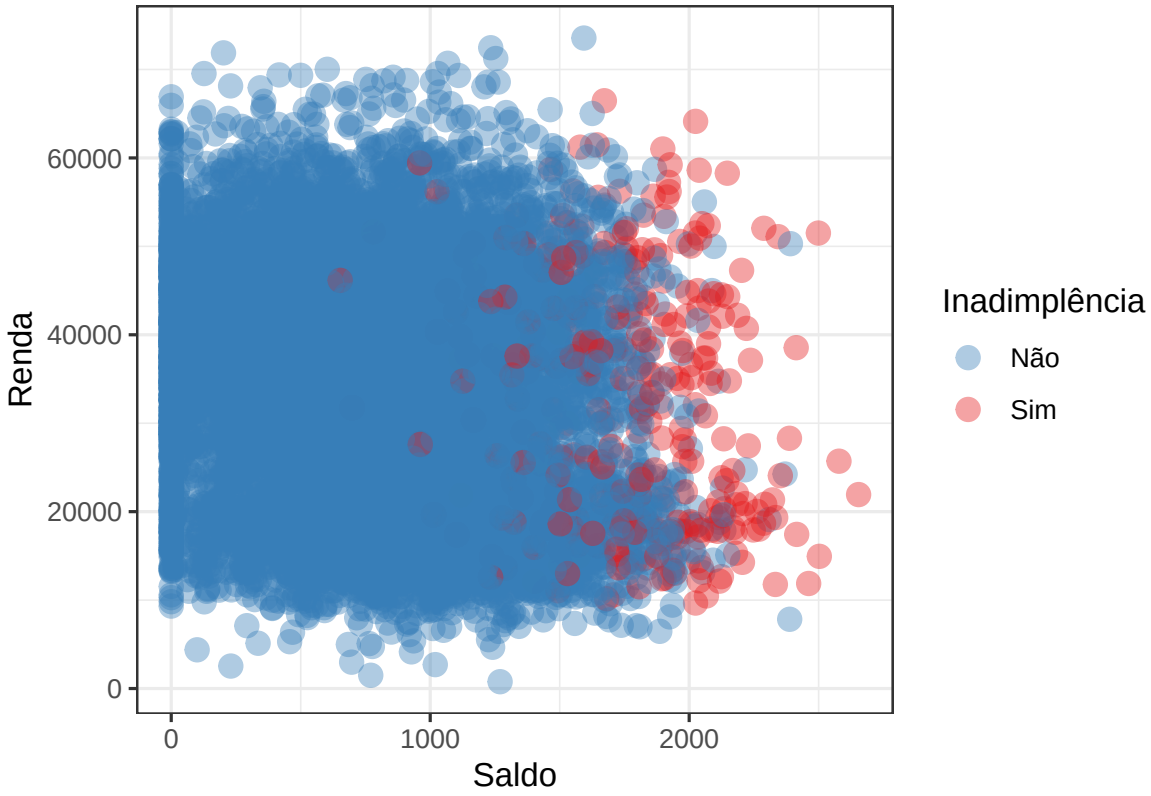
Figura 1: Distribuição da Renda e Saldo do cliente



Pelo gráfico acima podemos perceber que os clientes que estão ou não em situação de inadimplência tem uma distribuição de renda bastante parecida, tendo bastante interseção entre os dois níveis. No entanto, os clientes que não estão em situação de inadimplência, tem uma frequência de renda maior, o que faz sentido, já que esses clientes tem mais condições de arcar com as suas dívidas. Já na distribuição do saldo dos clientes, podemos ver que os clientes com saldos menores tendem a não não estar inadimplentes, enquanto os que estão em

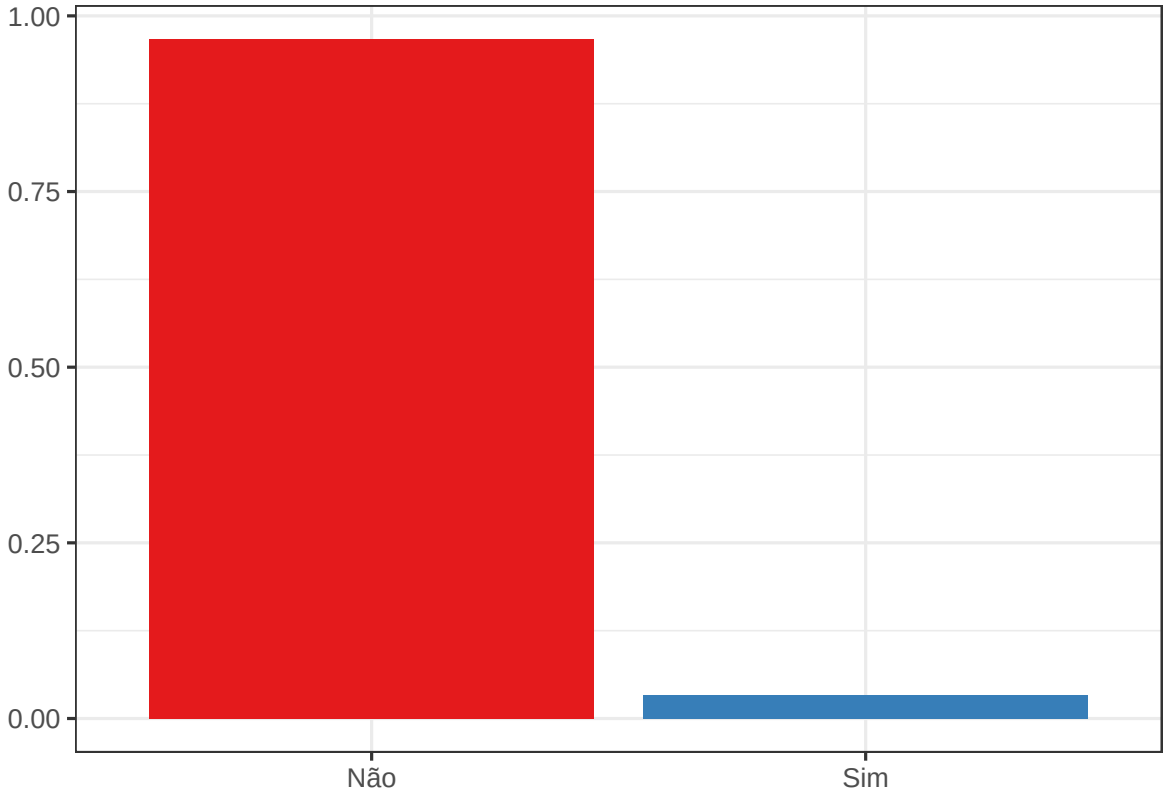
situação de inadimplência tem saldos maiores. Isto faz sentido, pois saldos mais altos podem indicar níveis mais elevados de dívida.

Figura 2: Renda em função do saldo do cliente



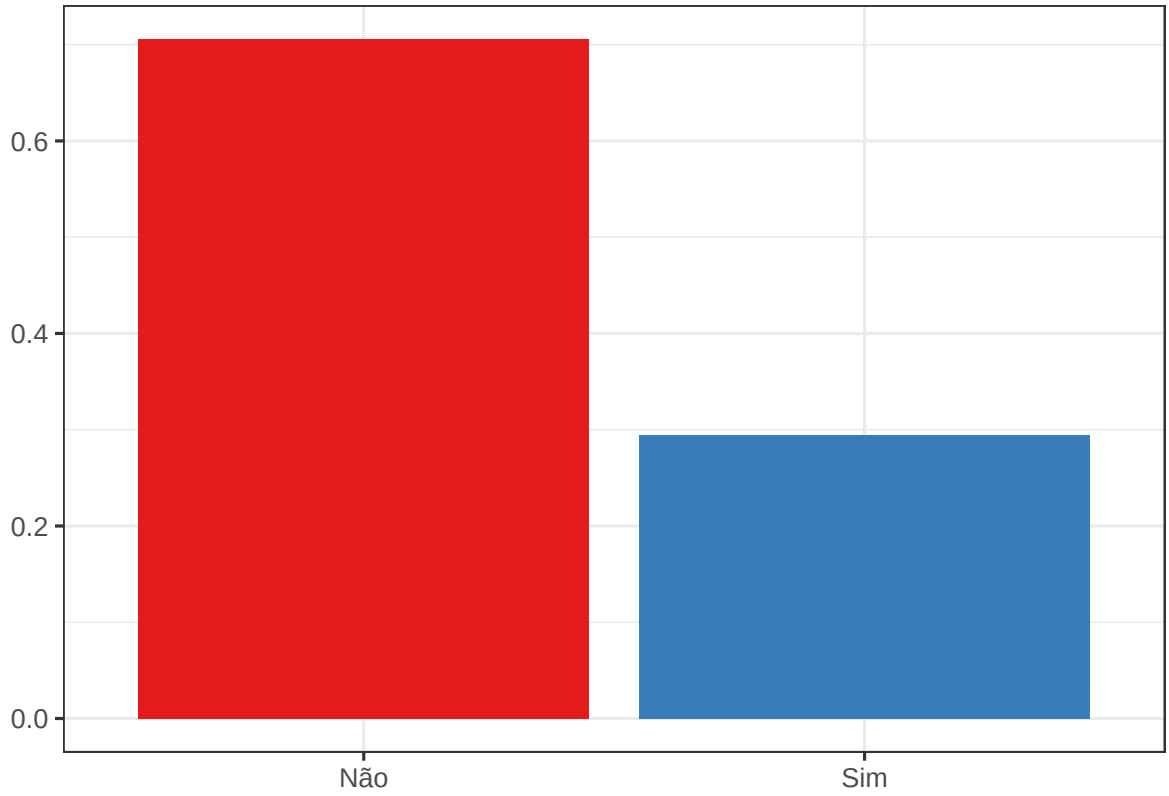
Assim como foi observado anteriormente no gráfico de densidade, a renda dos clientes não parecem indicar diferenças em situação de inadimplência. No entando, podemos observar novamente analizando apenas o eixo das abscissas, que saldos maiores parecem caracterizar aqueles clientes que costumam estar mais endividados.

Figura 3: Frequência dos clientes em inadimplência



O gráfico acima nos entrega a informação da porcentagem de clientes em situação de inadimplência. Dessa forma, podemos ver que 96,67% dos clientes não estão em situação de inadimplência. Isso pode significar que o modelo de regressão linear talvez fique melhor para classificar aqueles clientes que não estão em situação de inadimplência, pois a classe que está sendo modelada tem níveis desbalanceados.

Figura 4: Frequência dos clientes que são ou não estudantes



Agora observando a porcentagem porcentagem dos clientes que são estudantes, vemos que o banco de dados é composto por maioria de clientes não estudantes, com 70,56% não sendo estudantes.

Construção do modelo

Escolha das variáveis e adequação do modelo

O modelo foi inicialmente ajustado considerando todas as variáveis do banco de dados. Isto é, *Estudante*, *Saldo* e *Renda*. No entanto, após o ajuste do modelo foi contastado que uma das variáveis é não significativa para a regressão logística, tendo sido obtido a seguinte tabela para teste de significância dos coeficientes do modelo:

Tabela 1: Significância dos coeficientes de antigo modelo ajustado

	Coeficiente	Erro Padrão	Estatística	$Pr(>\|\|z\|\|)$
(Intercept)	-11.06	0.5502	-20.096	$2x10^{-16}$
Saldo	0.0057	$2.616x10^{4}$	22.132	$2x10^{-16}$
Renda	$6.526x10^{6}$	$9.037x10^{6}$	0.722	0.4702
Estudante_Sim	- 0.5244	0.2630	-1.994	0.0461

A partir da análise da tabela, podemos ver que apenas a variável *Renda* não é significativa para a classificação da situação de um cliente. Dessa forma, a variável renda foi removida e foi verificada novamente a significância de suas variáveis:

Tabela 2: Significância dos coeficientes de novo modelo ajustado

	Coeficiente	Erro Padrão	Estatística	$Pr(>\|\|z\|\|)$
(Intercept)	-10.80	0.4144	-26.06	$2x10^{-16}$
Saldo	0.0057	0.0002	22.16	$2x10^{-16}$
Estudante_Sim	-0.6720	0.1643	-4.09	$4.31x10^{-5}$

Com a remoção da variável renda, é possível ver que todas as variáveis são significativas. Assim, chegamos no seguinte modelo para regressão logística:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -10.6228 + 0.0056 Saldo_i - 0.8041 Estudante\_Sim_i$$

Agora, para verificar a adequabilidade do modelo, foi feito um teste de hipóteses ao nível de 5% de significância. Logo, temos a seguinte tabela:

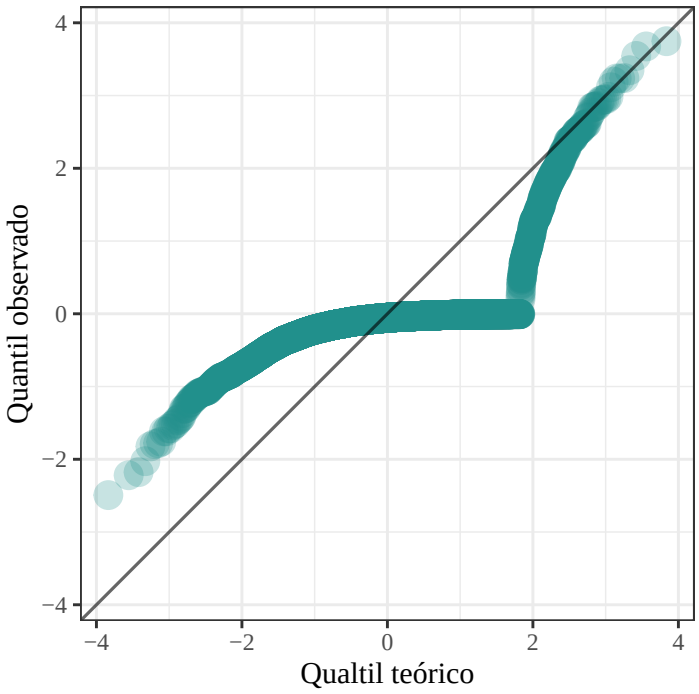
Tabela 3: Teste para adequabilidade do modelo

	Resultado Final
$Desvio/\phi$	1271.217
χ^2	1355.276

Portanto, como o desvio é menor que o quantil χ^2, o modelo é indicado para classificar a situação do cliente quanto ao cumprimento dos prazos.

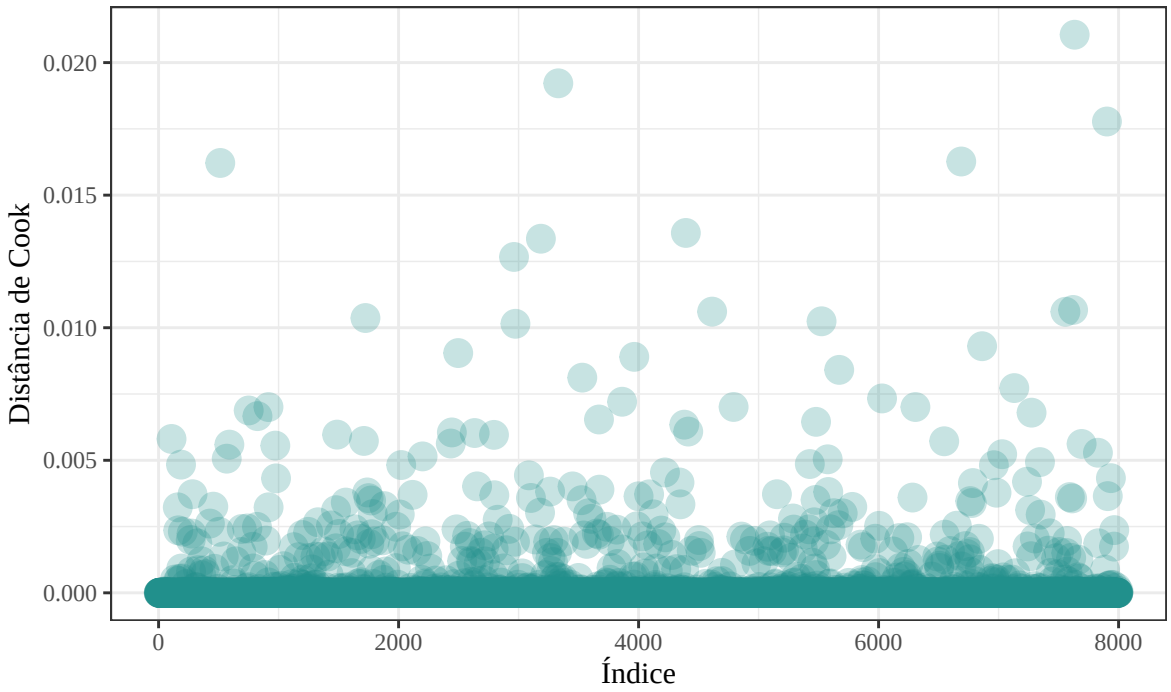
Análise de diagnósticos

Figura 5: Q-Q plot dos resíduos padronizados do modelo



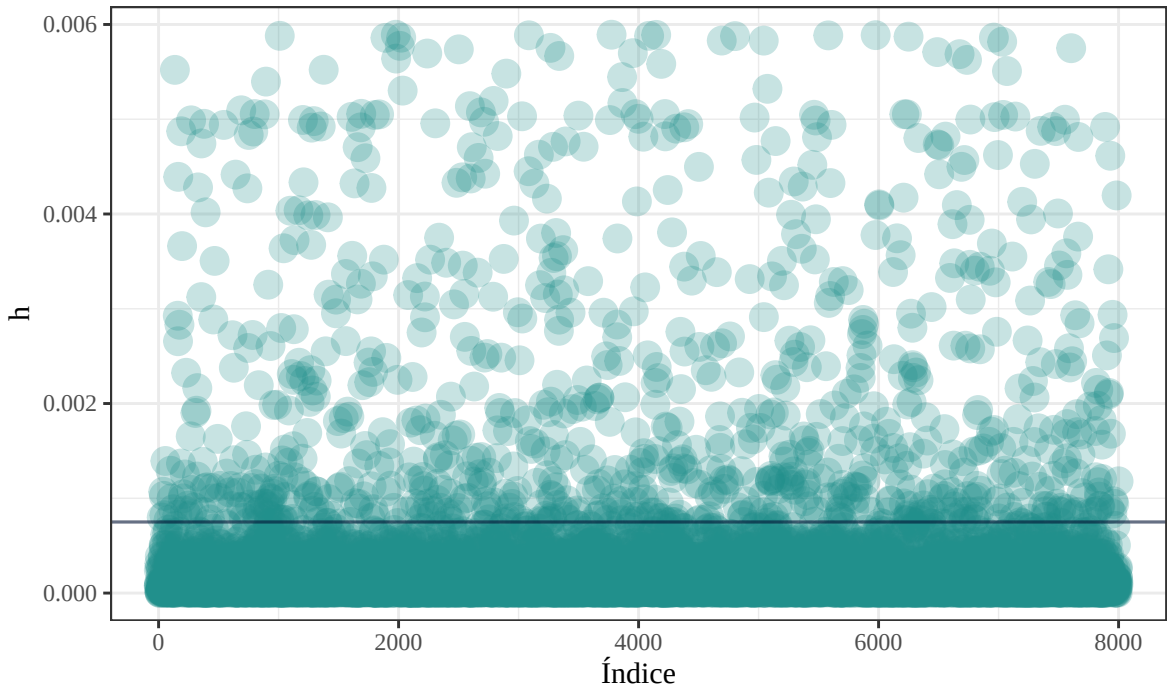
O gráfico de QQ-plot, que serve para verificar a normalidade dos resíduos do modelo, não parece indicar normalidade dos resíduos padronizados.

Figura 6: Pontos influentes



A partir da análise do gráfico com a distância de Cook, vemos que o modelo não possui pontos influentes.

Figura 7: Pontos de alavanca



Apesar do modelo não possuir pontos influentes, vemos que possui muitos pontos de alavanca, o que pode prejudicar ao fazer as predições. Além disso, o gráfico abaixo identifica os pontos aberrantes que podem estar presentes. É possível ver que há muitos pontos aberrantes, pois algumas observações ultrapassam aquele limite superior e inferior.

Figura 8: Pontos aberrantes

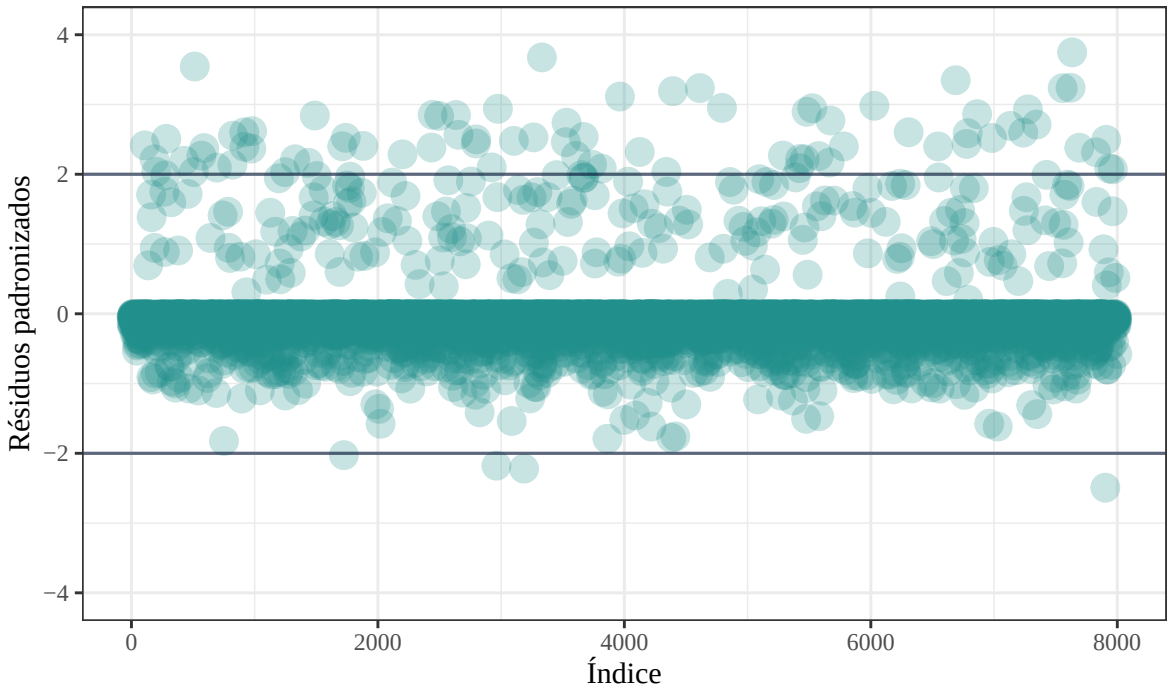
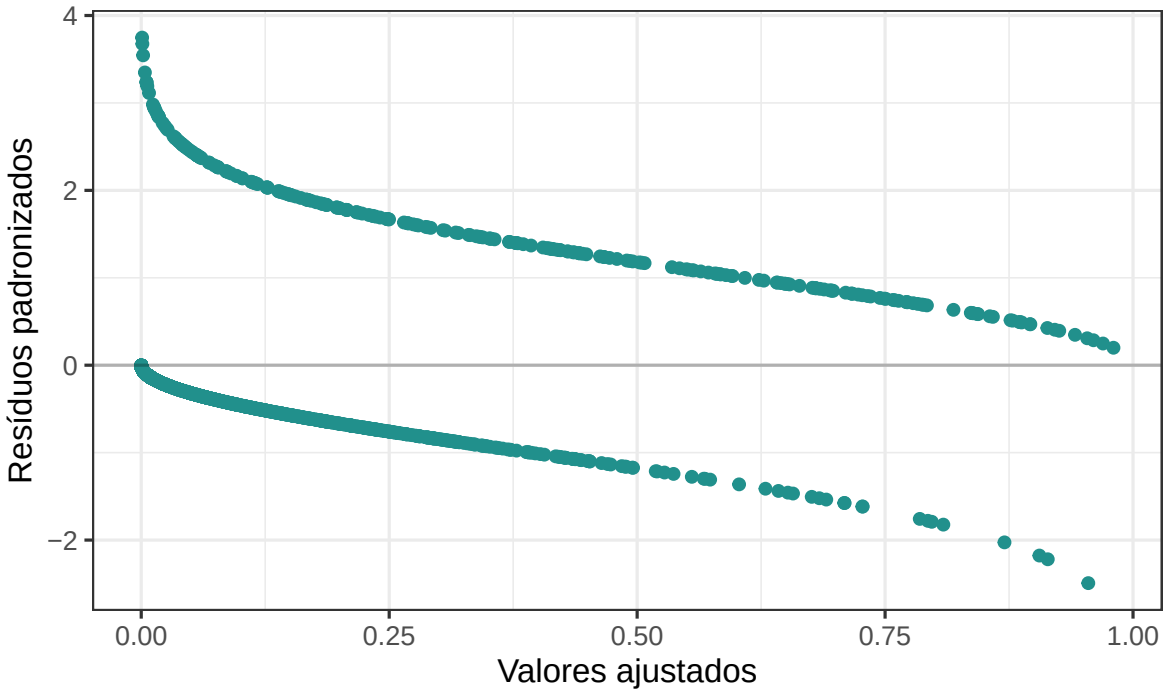


Figura 9: Resíduos padronizados versus valores ajustados



É difícil notar algum padrão de mudança de variância dos resíduos dependentes dos valores ajustados. Dessa forma, os resíduos parecem ser homocedástico.

Testes para normalidade e homocedasticidade

Foram realizados dois testes para a suposição de normalidade e homocedásticidade. Para suposição de normalidade, foram realizados os testes de Lilliefors e Anderson-Darling, já para a homocedásticidade foi o teste de Breusch-Godfrey e Goldfeld-Quandt. Assim como já foi identificado pelo qq-plot, vimos que a suposição de normalidade é rejeitada pelos dois testes escolhidos. Agora, para a homocedásticidade, os testes não rejeitam a suposição.

Teste	Suposiçao	$p - valor$
Lilliefors	Normalidade	2.2e-16
Anderson-Darling	Normalidade	2.2e-16
Breusch-Godfrey	Homocedasticidade	0.8126
Goldfeld-Quandt	Homocedasticidade	0.9985

Avaliação do modelo final

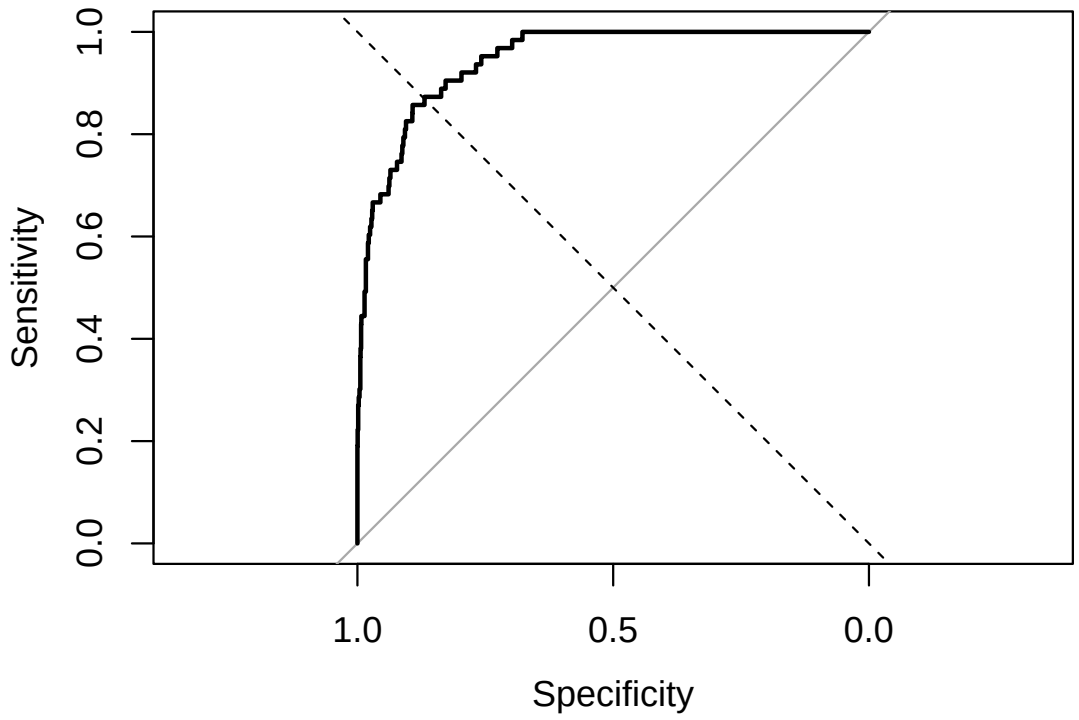
Tabela 5: Matriz de confusão

		Observados	
		Não	Sim
Previstos	Não	1923	39
	Sim	14	24

Pela matriz de confusão, vemos que 1923 dos clientes foram previstos corretamente como não sendo inadimplentes e 14 foram previstos como inadimplentes, mas não são. Já para os clientes inadimplentes os resultados ficaram um pouco piores, o que faz sentido já que a variável da situação dos clientes é bastante desbalanceada. Assim, podemos ver que 39 dos clientes inadimplentes foram incorretamente classificados como não sendo inadimplentes e 24 foram corretamente classificados como sendo inadimplentes. Ainda, o modelo final obtido teve uma acurácia de 97,35%, que é a proporção de classificações corretas da regressão logística. O modelo obteve uma sensibilidade de 99.28%, o que significa que 99.28% dos casos em que o cliente não é inadimplente foi corretamente classificado pelo modelo. Já para os clientes que são inadimplentes, apenas 38.10% foi classificado corretamente pelo modelo.

Observando a curva ROC, é possível ver que o modelo está classificando bem a situação de cada um dos clientes. Com a regressão logística final, foi possível chegar a um AUC de 94.64%.

Figura 10: Curva ROC da regressão logística ajustada



Interpretação da regressão logística

Pela tabela 2 em que foi avaliada a significância dos coeficientes do modelo, temos que o coeficiente da variável *Estudante* e *Saldo* foi -0.6720 e 0.0057, respectivamente. Analisando somente a variável *Saldo*, vemos que se mantendo constante todas as variáveis, as chances de um cliente ser inadimplente diminui em 0.43% a cada dolár de saldo. Agora analisando a variável dicotômica *Estudante*, podemos calcular a probabilidade de um cliente que não é estudante ser inadimplente. Dessa forma, podemos prosseguir da seguinte maneira: primeiro consideramos o caso em que o cliente é um estudante, depois fazemos o mesmo para o caso em que não é um estudante. Assim, fixamos o saldo e fazemos a razão desses dois cenários. Logo, chegamos que $\exp(0.941389) = 2.56354$ a chance de um cliente que não é estudante ser inadimplente é 156.35% vezes maior que um cliente que é estudante. O que faz sentido, pois clientes que não são estudantes tendem a ter rendas maiores e isso pode levar a saldos maiores que, pelo que foi visto na análise exploratória, é um dos maiores motivos de inadimplência.

Conclusões

A partir da análise exploratória foi possível identificar aquelas variáveis que são mais significantes para determinar a condição do cliente. Dessa forma, foi possível identificar que uma das únicas que representava alguma diferença foi o saldo, vimos que clientes com saldos maiores tem um risco maior de ser inadimplente. Foi possível confirmar também que os casos em que o cliente não era inadimplente, era melhor classificado pela regressão logística, o que pode ser explicado pelo desbalanceamento das classes da variável que está sendo modelada. Vimos que 96,67% dos clientes não eram inadimplentes, apenas 3,33% eram inadimplentes.

Foi possível também chegar a uma boa regressão linear para o caso em que os clientes não são inadimplentes. Observamos que o modelo final obteve uma acurácia de 97.35%, que é a proporção de classificações corretas. Não obstante, alcançou também uma sensibilidade de 99.28%, significando que a regressão logística obteve uma performance muito boa ao classificar os clientes não inadimplentes, pois 99.28% deles foram corretamente classificados. Agora considerando os clientes inadimplentes, apenas 38.10% foram corretamente classificados pela regressão logística.

Ainda, foi contastado que mesmo o modelo não tendo passado pela suposição de normalidade e ter muitos pontos de alavanca, foi possível obter boas classificações para o caso em que o cliente não é inadimplente.

Anexos