Classificação de Churn Utilizando um Modelo de Regressão Logística

Antonio C. da Silva Júnior
Orientador: Walmes M. Zeviani

Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em Data Science e Big Data

02 de outubro de 2020

Introdução

Retenção de clientes

- Atender um cliente se torna menos dispendioso a cada ano de relacionamento
- O custo para atrair um novo cliente pode ser de 5 a 25 vezes superior ao da manutenção de um já existente
- Desenvoler estratégias de retenção é uma prática comum em diversos segmentos
- Antever clientes propensos a abandonar o relacionamento (churn) se tornou um anseio constante

Proposta

- Um modelo preditivo para apoiar as estratégias de retenção da startup Olist
  - Classificação de churn
  - Interpretação dos principais motivos que impactam o desfecho
- Regressão logística
  - Modelagem híbrida multicritério considerando múltiplos decisores da empresa
  - Altamente confiável
  - Possibilita a interpretação direta dos parâmetros
  - Oferece a resposta na escala de probabilidade

Materiais e Métodos

Estruturação do conjunto de dados

- Extensivo processo de data wrangling
- Definição da variável resposta
- Criação de covariáveis de desempenho
- Inclusão de outras covariáveis qualitativas e quantitativas

Definição da variável resposta

- Vendedores pelo menos 30 dias inativos ($Y = 1$)
- Inatividade: não acesso à plataforma e a não realização de vendas online

Criação de covariáveis de desempenho

- Definida uma data de corte
- Período de 90 dias dividido em 2 subperíodos
- $V2/(V1 + V2)$, sendo V1 e V2 os valores calculados nos subperíodos
- Faturamento, ticket médio, pedidos cancelados, produtos publicados...

Tabela 1.: Definição da data de corte

Vendedor	Data de corte
Definido como churn	Última atividade
Em atividade normal	Realização da análise

Tabela 2.: Interpretação das métricas de desempenho

Valor	Desempenho
0,5	Mantido
> 0,5	Aumentado
< 0,5	Reduzido

Outras covariáveis

- Qualitativas: tipo de plano, região, estágio...
- Quantitativas: faturamento, pedidos, dias de atividade...
- Total: 31 covariáveis

Regressão Logística

- Probabilidade de ocorrência de um evento de interesse

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

$$\pi_i = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip})}}$$

Regressão Logística

- Estimação por máxima verossimilhança

$$L = \prod_{i=1}^{n} \left[\pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i} \right]$$

$$\log L = \sum_{i=1}^{n} \left\{ \left[Y_i \ln(\pi_i) \right] + \left[(1 - Y_i) \ln(1 - \pi_i) \right] \right\}$$

Ajuste do modelo e seleção de covariáveis

- Amostragem de 75% para o treinamento do modelo
  - 47,3% de ocorrência de *churn*
- Validação cruzada K-fold com 5 folds
- Modelo completo: todas as covariáveis
- Modelo restrito: algoritmo *stepwise*
  - Minimização do AIC ($-2 \log L + 2p$)
  - Múltiplo de penalização utilizado: 3,841459 (χ^2 com 1 grau de liberdade e 5% de significância)
  - P-valor = 0,05 como valor crítico para seleção das covariáveis

Resultados e discussões

Teste da razão da verossimilhança

- TRV $< \chi^2$ com 17 graus de liberdade e 5% de significância
- A qualidade do ajuste não foi afetada (H_0 não rejeitada)
- Optou-se pelo modelo restrito com 14 covariáveis

$$\text{TRV} = -2(\log L_{\text{completo}} - \log L_{\text{restrito}})$$

Covariáveis selecionadas

Tabela 3.: Covariáveis utilizadas pelo modelo

Covariável	Descrição	Suporte
Métricas de desempenho		
X2	Desempenho do número de pedidos cancelados ou suspensos no período	$[0, 1]$
X4	Desempenho da média de dias de atraso (postagens) no período	$[0, 1]$
X5	Desempenho da média de dias de atraso (entregas) no período	$[0, 1]$
X6	Desempenho do faturamento no período	$[0, 1]$
X9	Desempenho do ticket médio no período	$[0, 1]$
Qualitativas		
X11	Necessita de dias adicionais para postagem	$\{0, 1\}$
X13	Está no estágio I	$\{0, 1\}$
X14	Está no estágio R	$\{0, 1\}$
X20	Localizado na região Sudeste	$\{0, 1\}$
Quantitativas		
X22	Valor total do faturamento	$\mathbb{R}_+$
X27	Número total de produtos publicados	$\mathbb{N}$
X29	Número de dias desde a contratação até a primeira venda	$\mathbb{N}$
X30	Número de dias desde a contratação até estar pronto para operar	$\mathbb{N}$
X31	Número de dias em atividade no período	$\mathbb{N}$

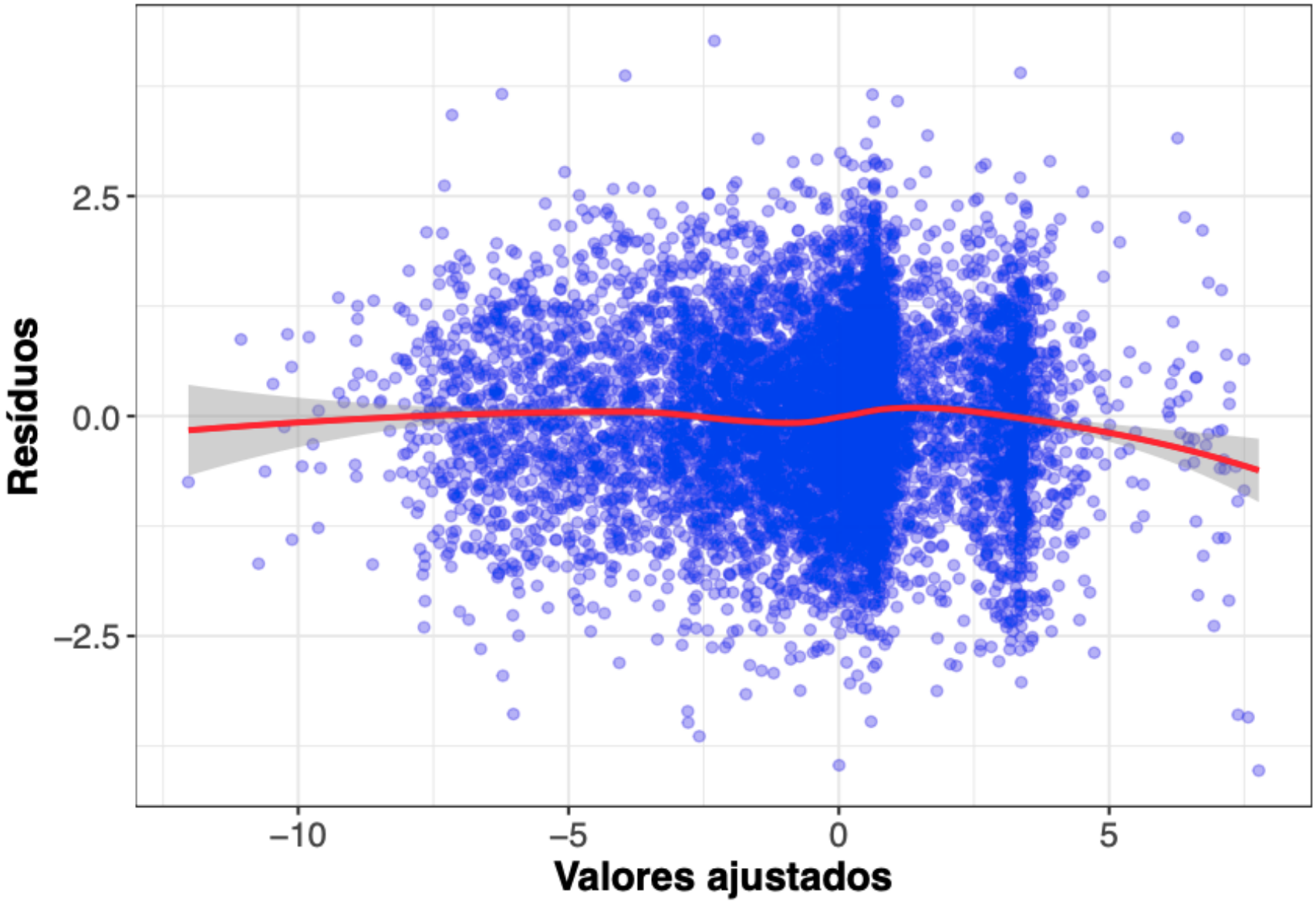
Análise dos resíduos

- Resíduos quantílicos aleatorizados
  - $r_i = \phi^{-1}(u_i)$
  - ϕ^{-1}: inversa da função de distribuição acumulada da normal padrão
  - $u_i = F(y_i; \mu_i, \phi)$: com base na distribuição acumulada do modelo proposto
  - Normalmente distribuídos e com variância constante

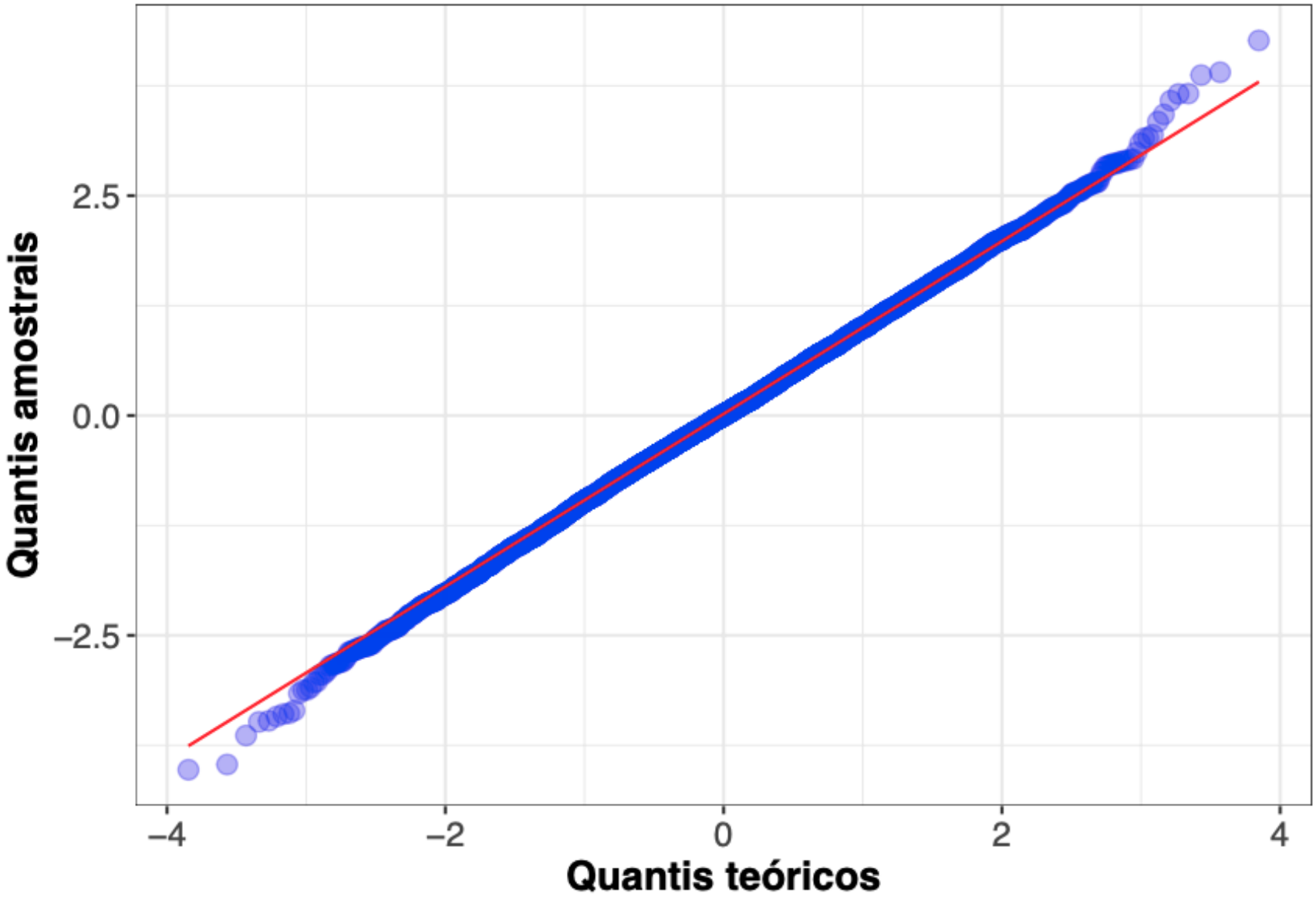
Análise dos resíduos

- Variabilidade constante, centrados predominantemente entre -2 e 2
- Aderentes à distribuição normal
- Não houve violação dos pressupostos

Análise dos resíduos



Análise dos resíduos



Interpretação das estimativas dos parâmetros

Tabela 4.: Estimativas dos parâmetros do modelo

Covariável	Estimativa	Erro padrão	Wald	P-valor
Intercepto	1.5513	0.1316	11.7902	0.0000
X2	0.3693	0.0387	9.5437	0.0000
X4	-0.3558	0.0444	-8.0085	0.0000
X5	-0.2567	0.0487	-5.2749	0.0000
X6	-0.8296	0.1052	-7.8829	0.0000
X9	0.4532	0.1006	4.5028	0.0000
X11	-0.3950	0.0697	-5.6703	0.0000
X13	-2.6981	0.1237	-21.8040	0.0000
X14	-2.2880	0.1246	-18.3608	0.0000
X20	0.1598	0.0628	2.5453	0.0109
X22	-0.1925	0.0629	-3.0595	0.0022
X27	-0.1717	0.0748	-2.2949	0.0217
X29	-0.4516	0.0321	-14.0894	0.0000
X30	0.4201	0.0457	9.1935	0.0000
X31	-1.6460	0.0863	-19.0675	0.0000

Interpretação das estimativas dos parâmetros

- Fatores que mais impactam na redução da chance de *churn*:
  - Estar no estágio I (X13) ou no estágio R (X14)
  - Aumento do número de dias em atividade no perído (X31)
  - Melhora do desempenho do faturamento (X6)

Interpretação das estimativas dos parâmetros

Tabela 5.: Chances de ocorrência de churn para covariáveis quantitativas e qualitativas

Covariável	Estimativa	Chance	Variação (%)
X13	-2.6981	0.0673	-93
X14	-2.2880	0.1015	-90
X31	-1.6460	0.1928	-81
X30	0.4201	1.5221	52
X29	-0.4516	0.6366	-36
X11	-0.3950	0.6737	-33
X22	-0.1925	0.8249	-18
X20	0.1598	1.1733	17
X27	-0.1717	0.8422	-16

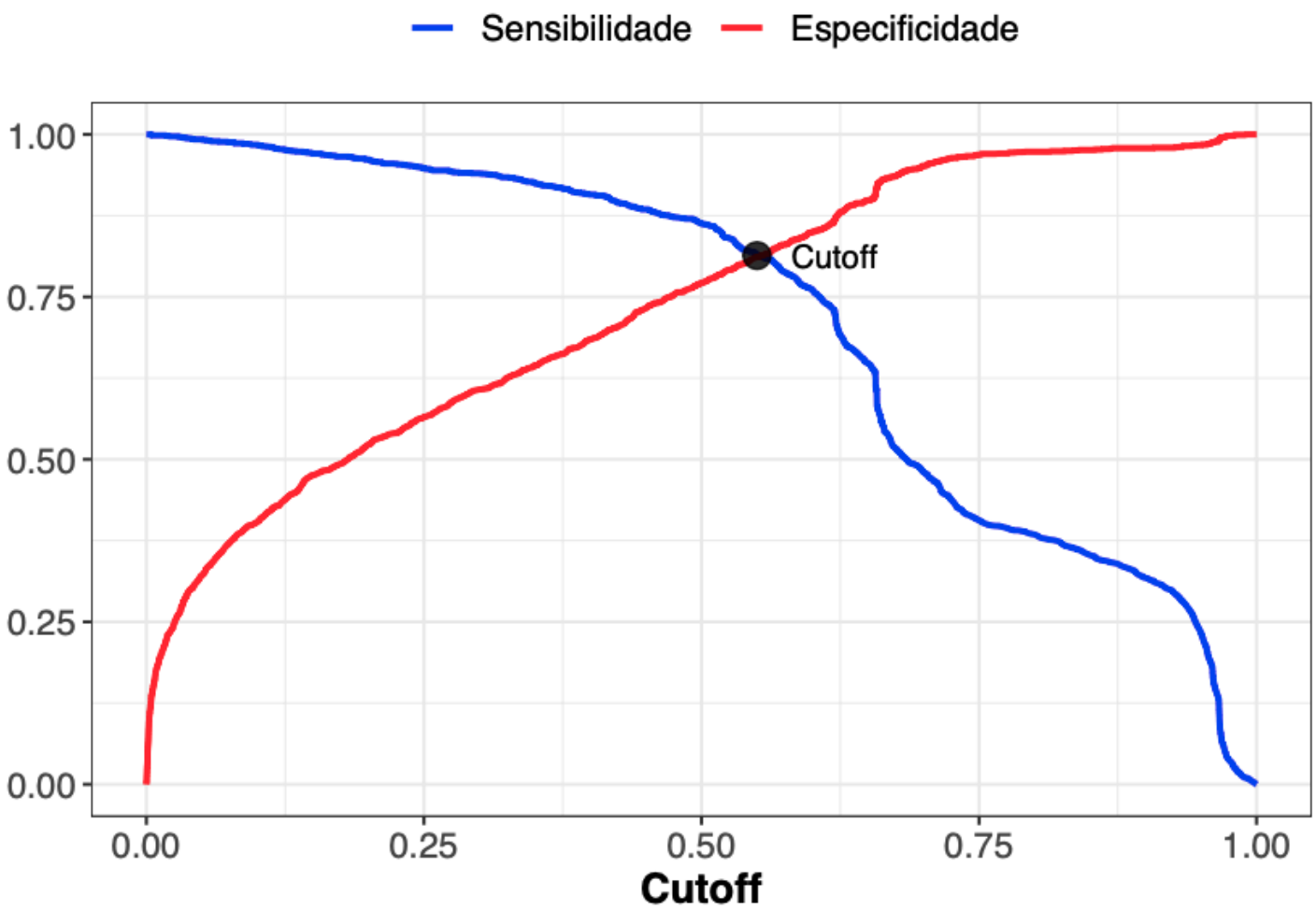
Interpretação das estimativas dos parâmetros

Tabela 6.: Chances de ocorrência de churn para covariáveis de desempenho

Covariável	Estimativa	Chance (0,5)	Variação (%)
X6	-0.8296	0.6605	-34
X9	0.4532	1.2543	25
X2	0.3693	1.2028	20
X4	-0.3558	0.8370	-16
X5	-0.2567	0.8795	-12

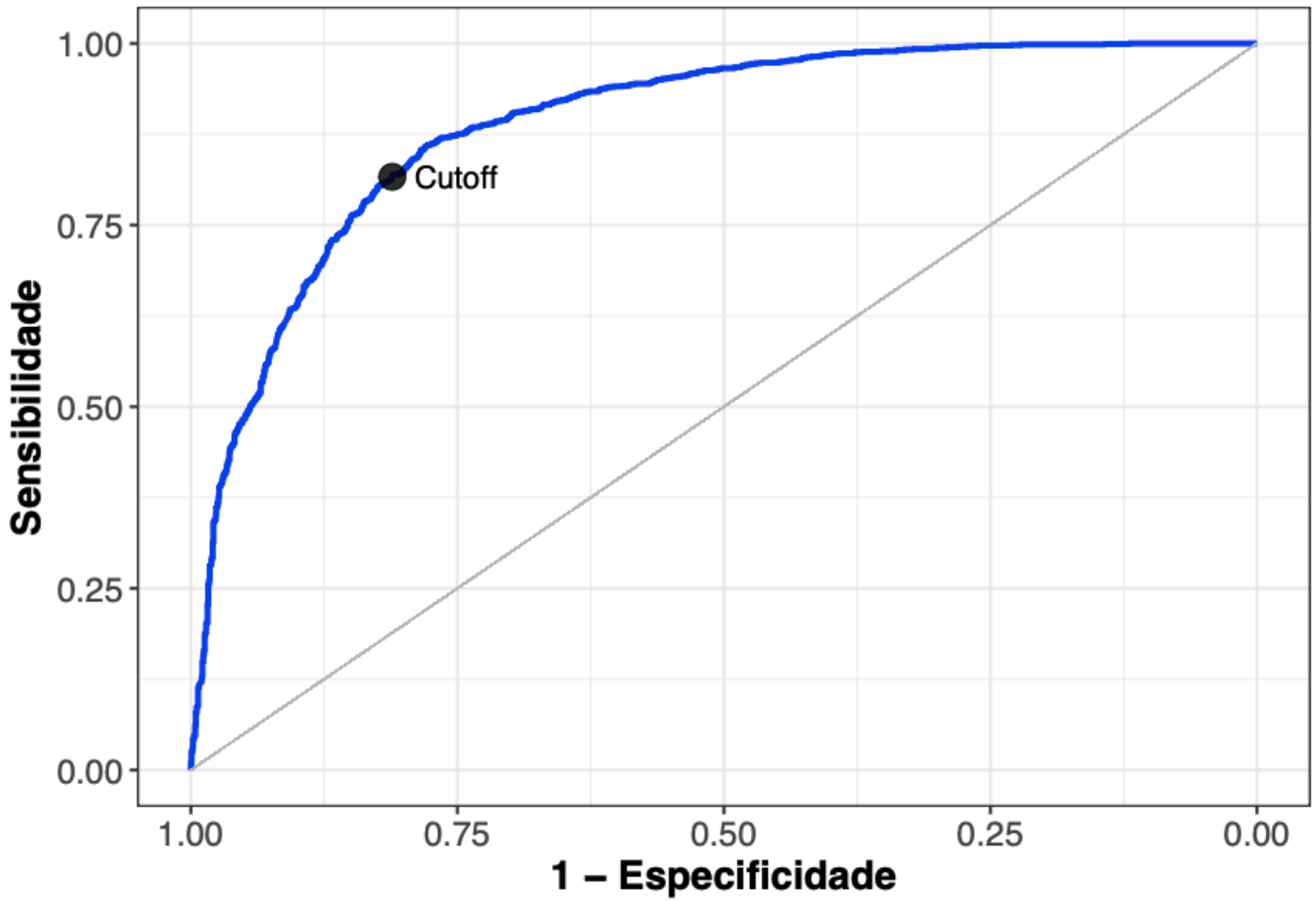
Escolha do cutoff e avaliação da curva ROC

- Cutoff: 0,55



Escolha do cutoff e avaliação da curva ROC

- AUC: 0,89



Avaliação do poder preditivo do modelo

Tabela 7.: Matriz de confusão

	Observado	
Predito	0	1
0	Verdadeiro negativo (VN)	Falso negativo (FN)
1	Falso positivo (FP)	Verdadeiro positivo (VP)

- Sensibilidade $= VP/(VP + FN)$: 0,82
- Especificidade $= VN/(VN + FP)$: 0,81
- Acurácia $= VN + VP/(VN + VP + FN + FP)$: 0,81 (IC de 0,80 até 0,83)

Conclusões

Considerações finais

- O modelo proposto equilibra poder de predição e interpretabilidade
- Oferece insights para ações de marketing personalizadas com foco na retenção de clientes
- A abordagem para a definição da variável resposta e criação das métricas de desempenho demonstrou-se eficaz

Obrigado.