



Classificação de Churn com Regressão Logística

Antonio C. da Silva Júnior

14/02/2021

A decorative graphic on the left side of the slide consisting of two overlapping squares. The top square is a lighter blue and the bottom square is a darker blue.

Objetivo

Apresentar o trabalho de conclusão do curso de Especialização em Data Science e Big Data da Universidade Federal do Paraná, intitulado **CLASSIFICAÇÃO DE CHURN UTILIZANDO UM MODELO DE REGRESSÃO LOGÍSTICA** (Silva Júnior, 2020)

Download da apresentação

- https://acsjunior.com/presentations/churn_cpo.pdf



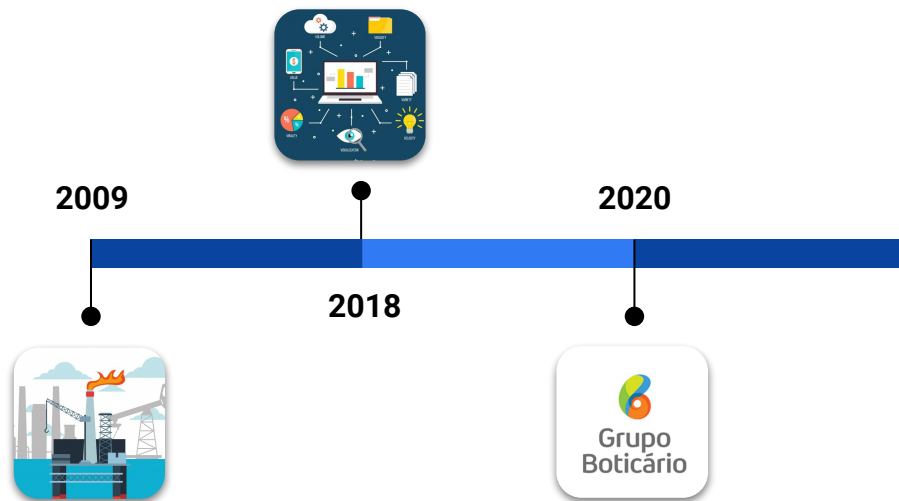
Quem sou eu?



**Antonio C. da Silva
Júnior**

Cientista de dados

- M.e Métodos Numéricos em Engenharia (UFPR, em andamento)
- Esp. Data Science e Big Data (UFPR, 2020)
- Tecg. Análise e Desenvolvimento de Sistemas (UNIP, 2015)





01

Introdução

Retenção de clientes

- O custo para adquirir um novo cliente pode ser de 5 a 25 vezes superior ao da manutenção de um cliente já existente
- Atender um cliente se torna menos dispendioso a cada ano adicional de relacionamento
- A retenção de clientes é essencial para o sucesso das empresas
- Desenvolvimento de estratégias de retenção se tornou uma prática comum entre empresas de diversos segmentos

A proposta

- Antever clientes propensos a abandonar o relacionamento
- Apoiar as estratégias de retenção de clientes da companhia
- Um modelo para classificação churn de que permita a interpretação dos principais motivos que impactam o desfecho

Por que regressão logística?

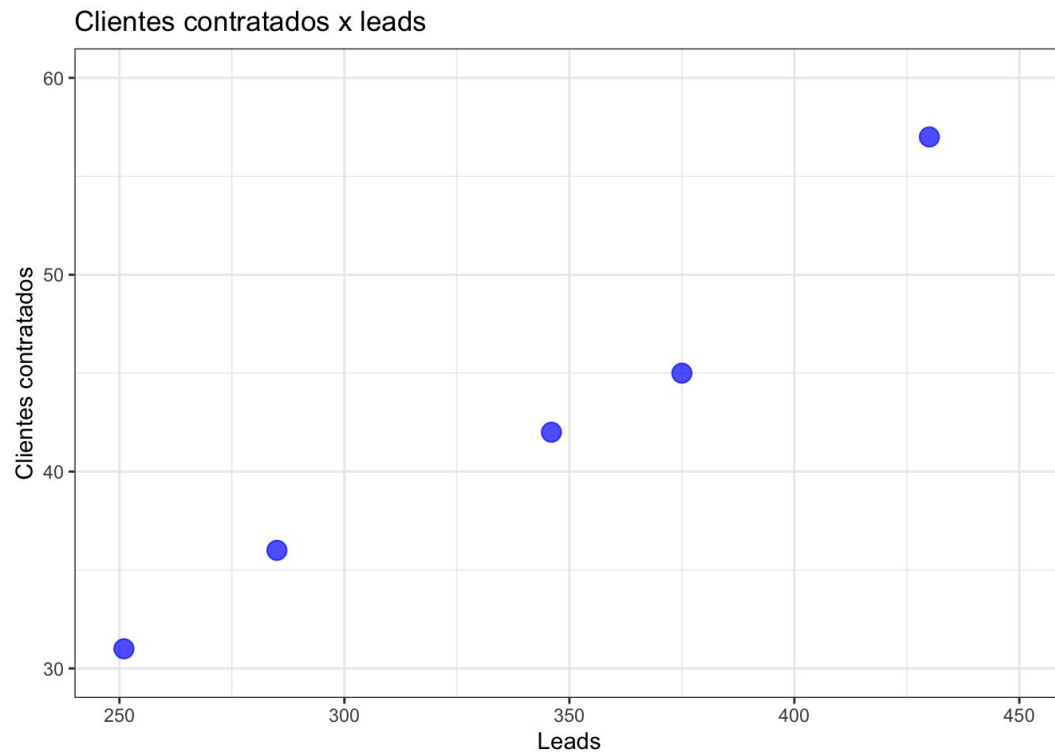
- Escolhida através de uma modelagem híbrida multicritério considerando múltiplos decisores da companhia
 - Métodos VIKOR + SAPEVO-M
 - <http://admpg.com.br/2020/anais/>
- Altamente confiável
- Possibilita a interpretação direta dos parâmetros
- Oferece a resposta na escala de probabilidade



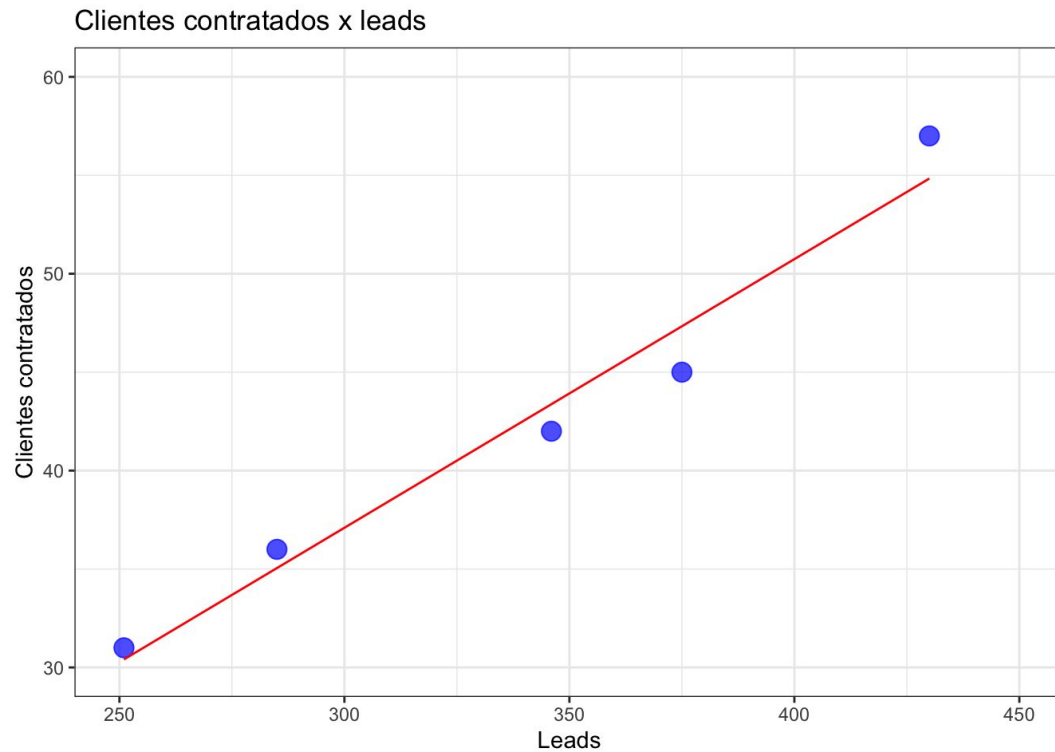
02

Regressão Logística

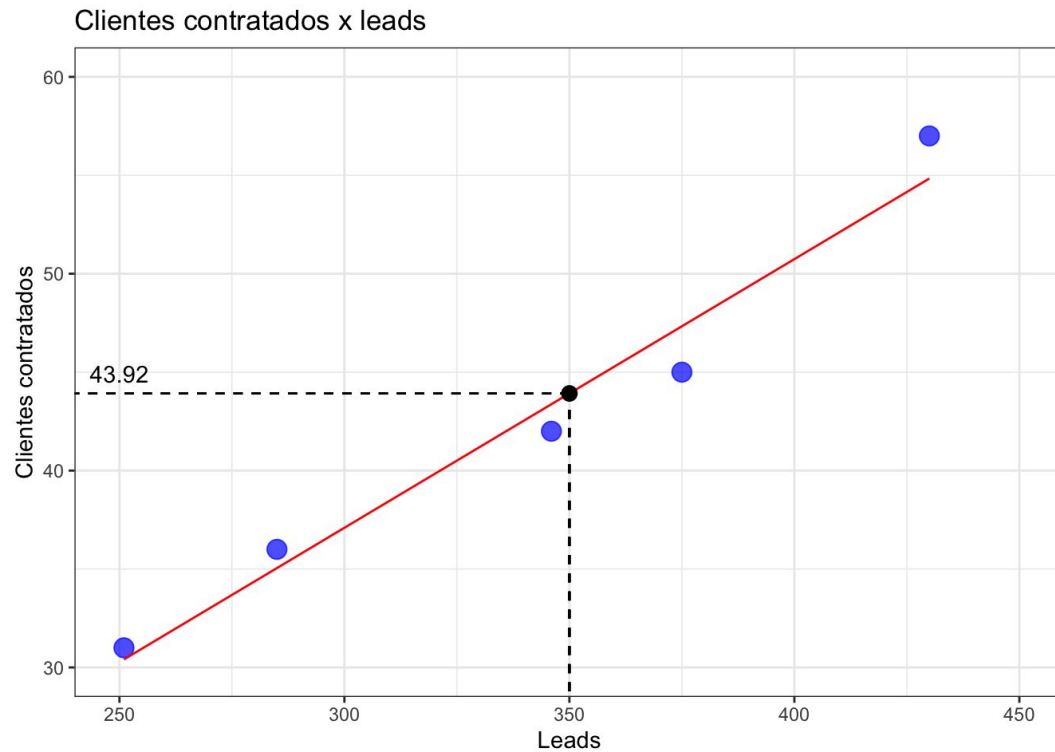
Regressão Linear



Regressão Linear

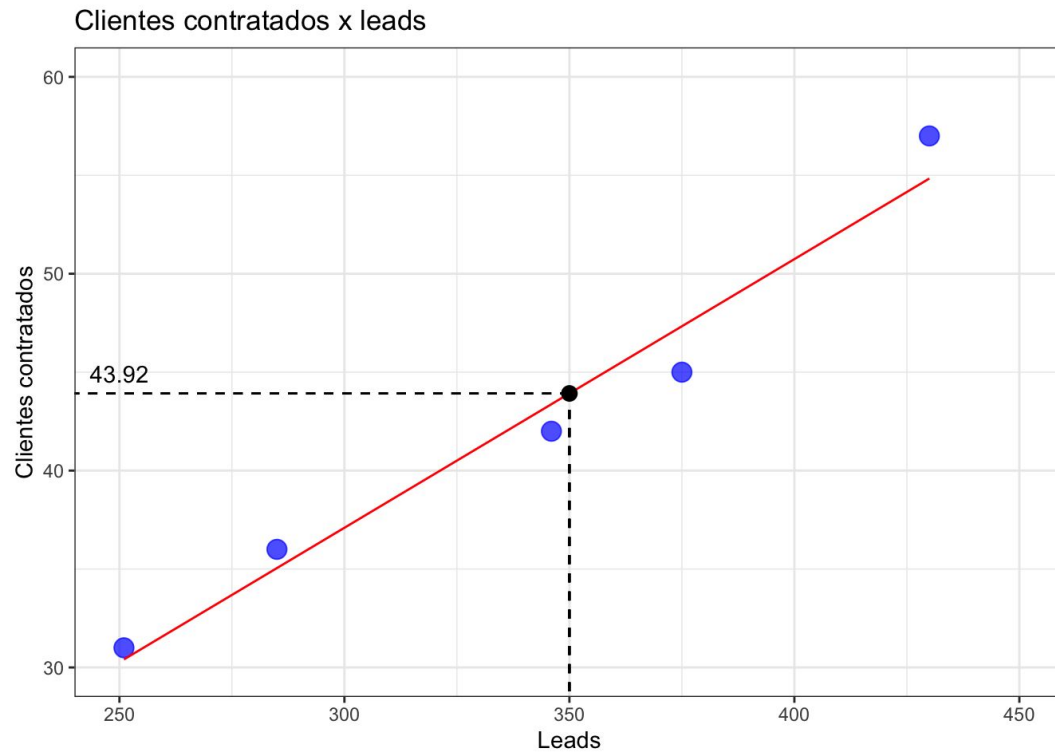


Regressão Linear



Regressão Linear

$$43,92 = \hat{\beta}_0 + \hat{\beta}_1 350$$

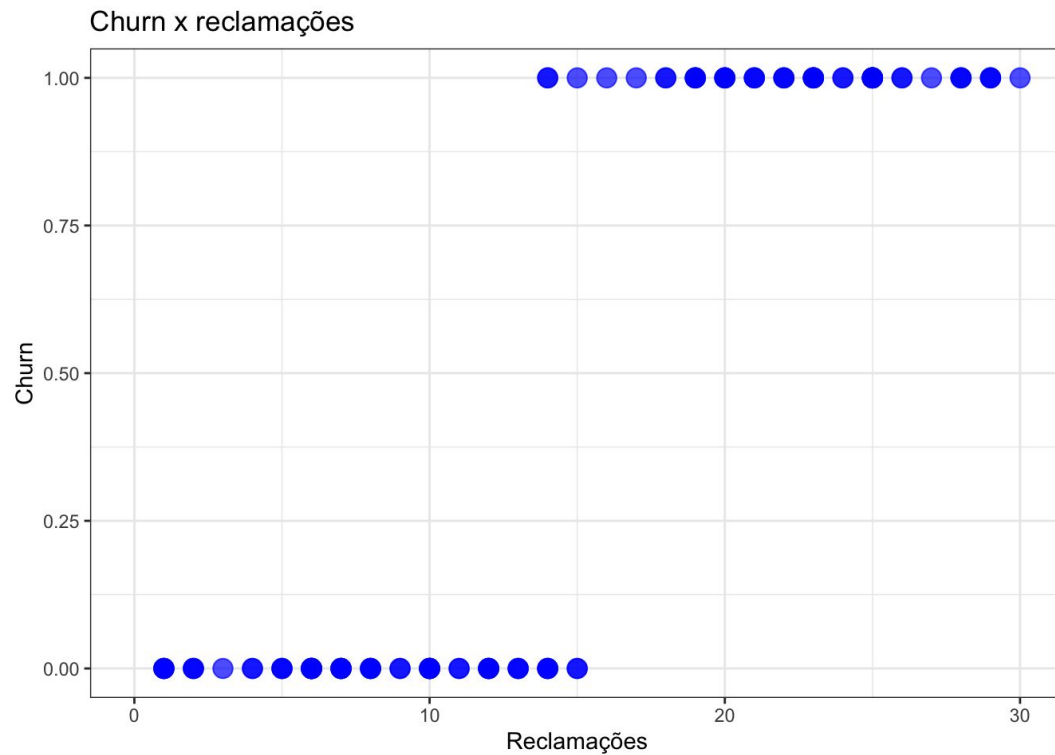


Modelos de regressão

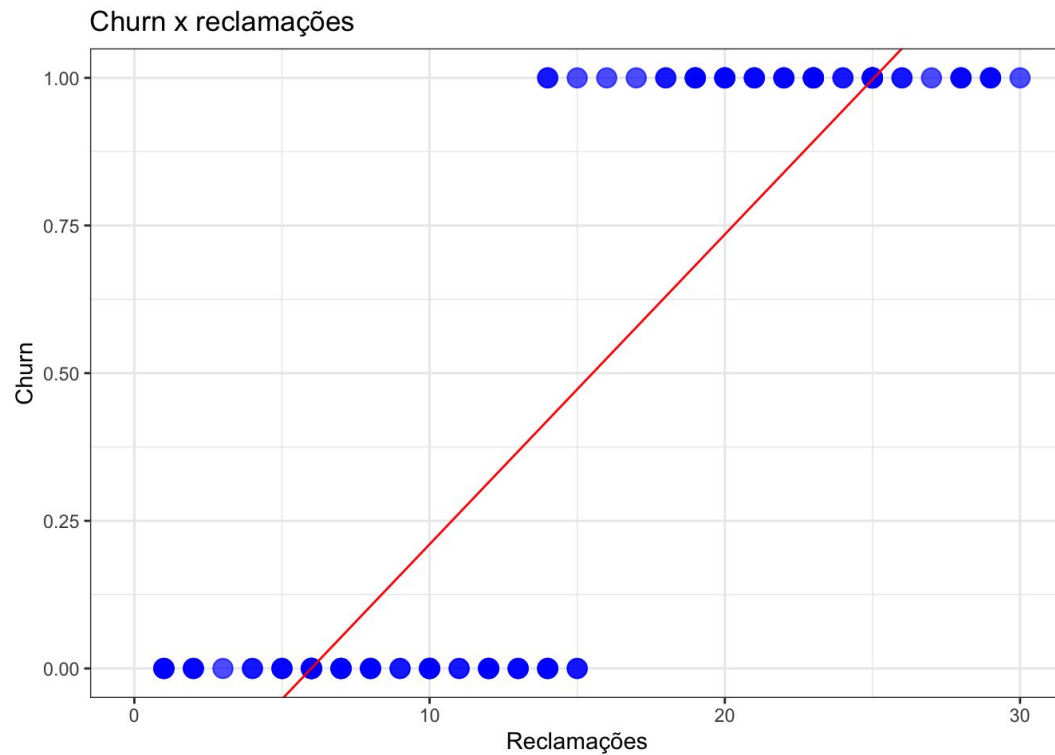
- Modelar a relação entre uma variável resposta y e um conjunto de covariáveis (x_1, \dots, x_n)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

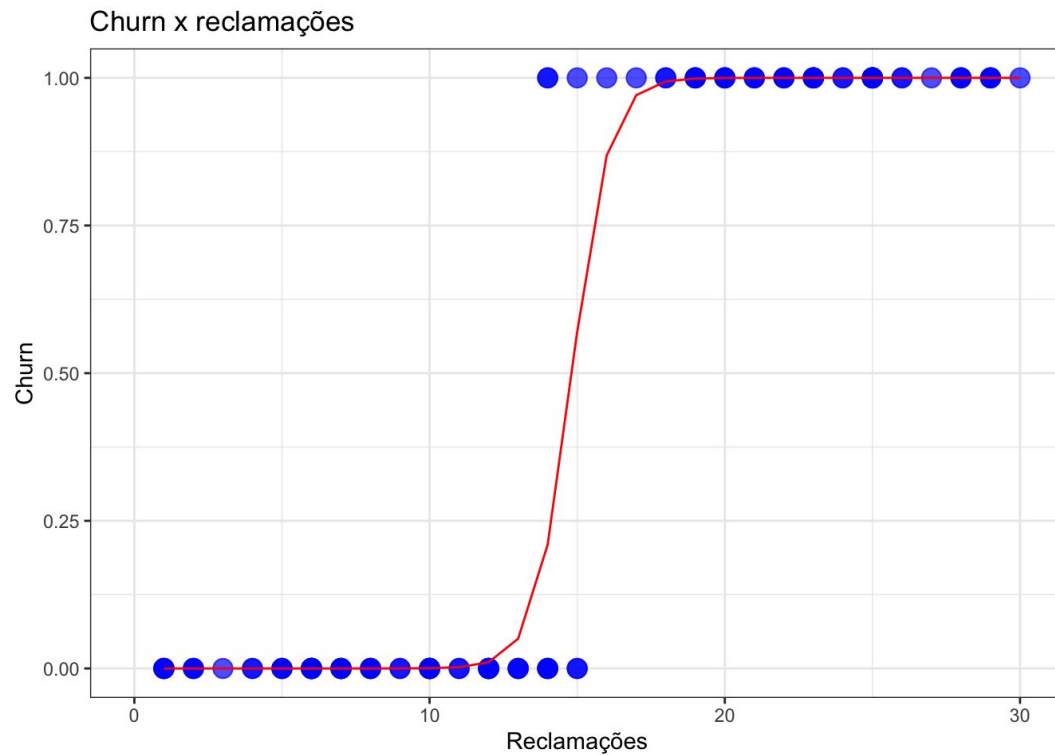
Regressão Logística



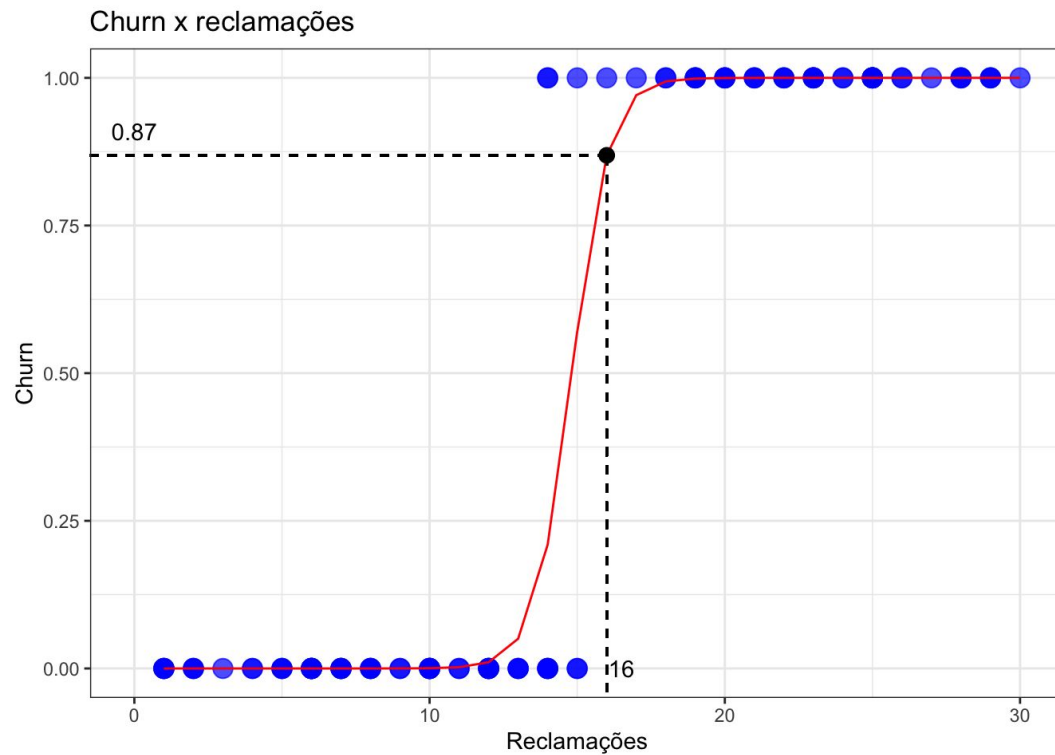
Regressão Logística



Regressão Logística

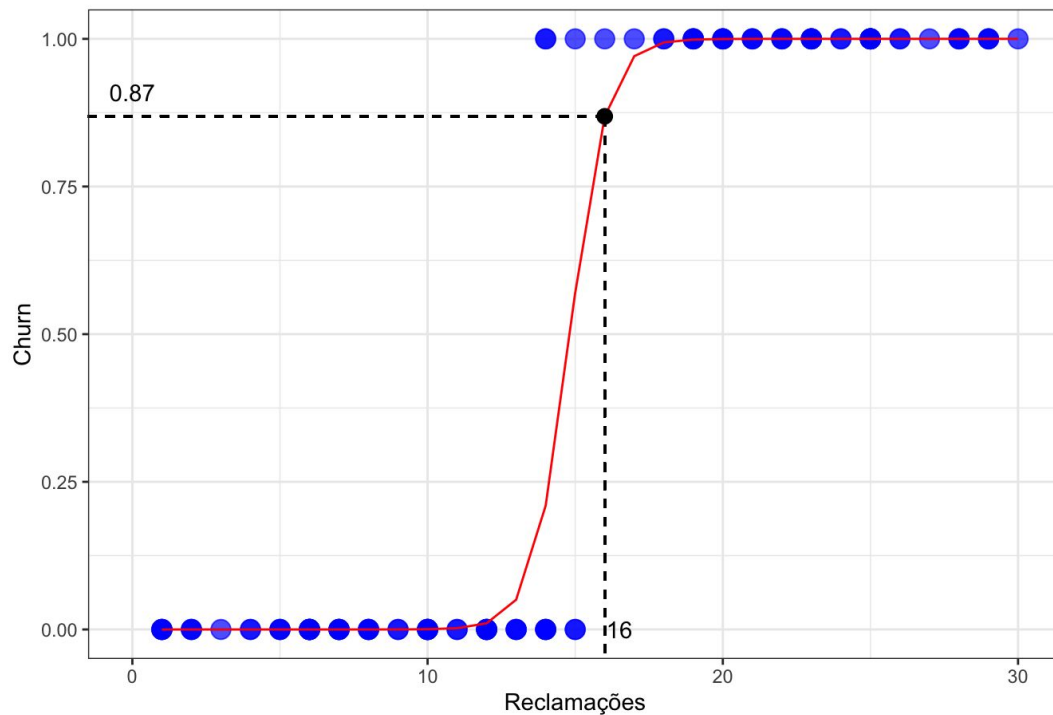


Regressão Logística



Regressão Logística

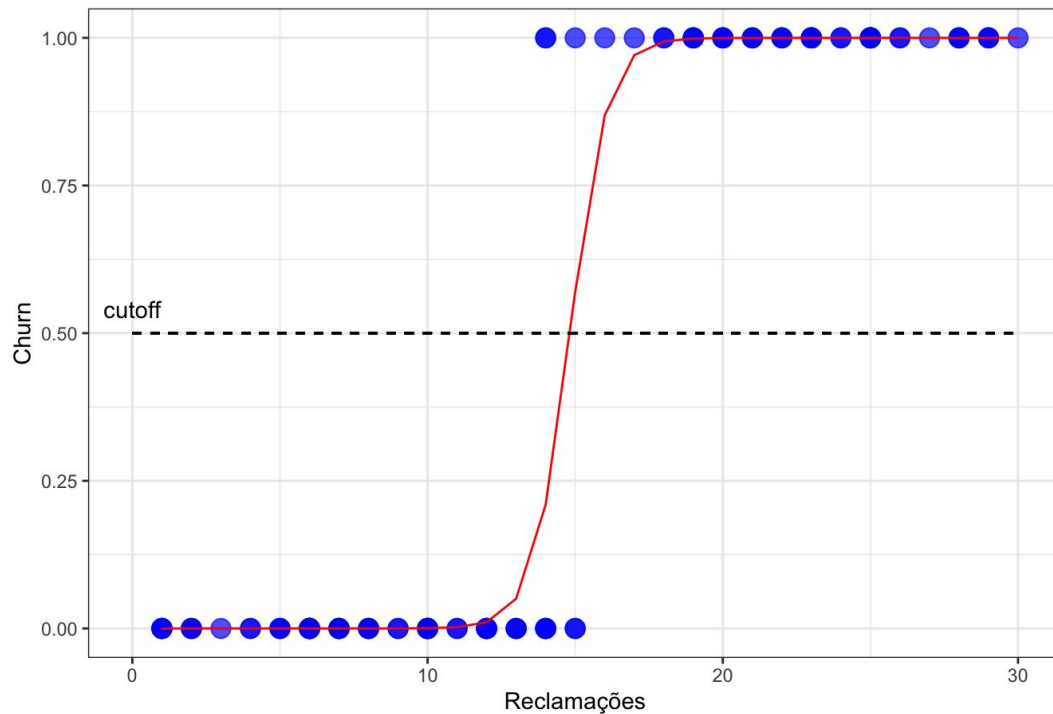
Churn x reclamações



$$0,87 = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 16}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 16}}$$

Regressão Logística

Churn x reclamações



$$0,87 = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 16}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 16}}$$

Modelos Lineares Generalizados

- Extensões dos modelos lineares clássicos
- Permite modelar variáveis resposta com outras distribuições da família exponencial de distribuições:
 - Binomial (dados binários) - ex. concessão de crédito
 - Poisson (dados de contagem) - ex. núm. casos de COVID19 em uma região
- Gama (dados contínuos e assimétricos) - ex. valor de imóveis

Modelos Lineares Generalizados

- Extensões dos modelos lineares clássicos
- Permite modelar variáveis resposta com outras distribuições da família exponencial de distribuições:
 - Binomial (dados binários) - ex. concessão de crédito
 - Poisson (dados de contagem) - ex. núm. casos de COVID19 em uma região
- Gama (dados contínuos e assimétricos) - ex. valor de imóveis

Modelos Lineares Generalizados

- Componente sistemático: $n_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$
 - Preditor linear do modelo
- Componente aleatório: y_1, y_2, \dots, y_n
 - Variáveis aleatórias independentes
- Função de ligação: $g(\pi_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$
 - Linearizar a relação entre os componentes aleatório e sistemático

Modelo de Regressão Logística

- A probabilidade de um particular cliente deixar a empresa é de 25%. Qual é a chance deste cliente deixar a empresa?

Modelo de Regressão Logística

- A probabilidade de um particular cliente deixar a empresa é de 25%. Qual é a chance deste cliente deixar a empresa?

$$\text{chance (odds)} = \frac{\pi_i}{1 - \pi_i} = \frac{0,25}{1 - 0,25} = \frac{1}{3}$$

Modelo de Regressão Logística

- Função de ligação: **logito** (baseada na distribuição logística)

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

Modelo de Regressão Logística

- Função de ligação: **logito** (baseada na distribuição logística)

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

$$\frac{\pi_i}{1 - \pi_i} = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

Modelo de Regressão Logística

- Função de ligação: **logito** (baseada na distribuição logística)

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

$$\frac{\pi_i}{1 - \pi_i} = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

Modelo de Regressão Logística

- Estimação dos parâmetros por máxima verossimilhança

$$p(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

Modelo de Regressão Logística

- Estimação dos parâmetros por máxima verossimilhança

$$p(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

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

Modelo de Regressão Logística

- Estimação dos parâmetros por máxima verossimilhança

$$p(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

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

$$LL = \sum_{i=1}^n [y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)]$$



03

Estruturação dos dados

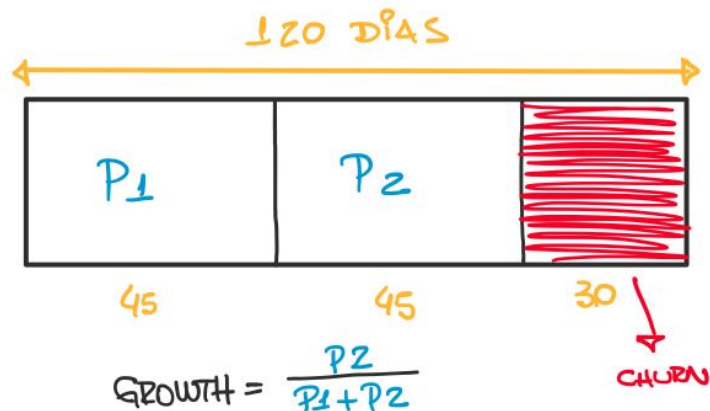
Definição da variável resposta

- Extensivo processo de data wrangling
- Clientes inativos por 30 dias consecutivos
- Criação de uma data de corte:

Cliente	Data de corte
Inatividade \geq 30 dias	Última atividade
Inatividade $<$ 30 dias	Realização da análise

Métricas de desempenho

- Mantidos no dataset somente os clientes com pelo menos 90 dias de histórico
- Dividido o período de 90 dias em 2 subperíodos



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

Outras covariáveis

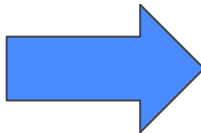
- Adição de outras covariáveis qualitativas e quantitativas
- Transformação das covariáveis qualitativas em dummies

Cliente	Plano contratado
1	A
2	B
3	B
4	C

Outras covariáveis

- Adição de outras covariáveis qualitativas e quantitativas
- Transformação das covariáveis qualitativas em dummies

Cliente	Plano contratado
1	A
2	B
3	B
4	C



Cliente	Plano B	Plano C
1	0	0
2	1	0
3	1	0
4	0	1

Conjunto de dados final

- Qualitativas
 - Tipo de plano contratado
 - Região
 - etc...
- Quantitativas:
 - Faturamento
 - Pedidos
 - Métricas
 - etc...
- Total: 31 covariáveis



04

Ajuste do modelo

Etapas

- Treino por validação cruzada k-fold (5 folds)
- Modelo completo (todas as covariáveis)
- Modelo restrito (algoritmo stepwise)
- Teste da razão da verossimilhança

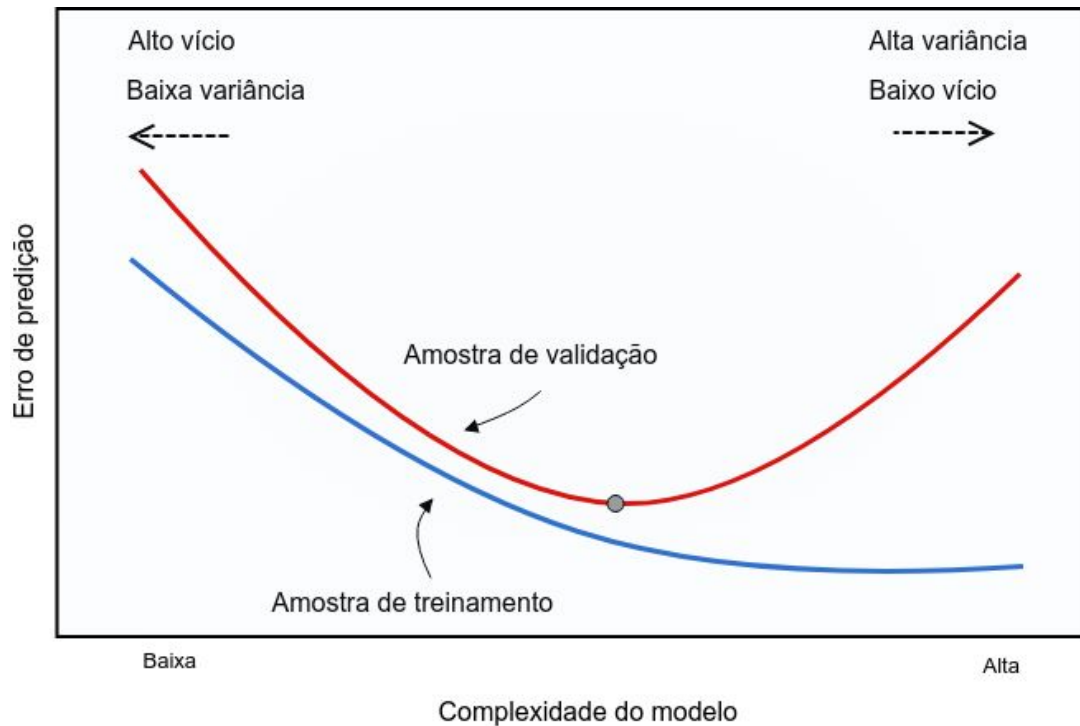
Algoritmo Stepwise

1. Ajusta-se o modelo com todas as p covariáveis
2. Avalia-se a exclusão e a inclusão de cada covariável (AIC)
3. Inclui-se (ou exclui-se) a covariável cuja inclusão (ou exclusão) resulta no menor AIC
4. Repete-se os passos anteriores até nenhuma inclusão (ou exclusão) resultar em um modelo com menor AIC

Critério de seleção

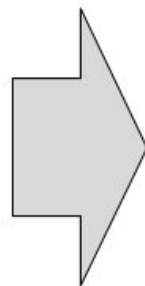
- Critério de informação de Akaike (AIC): $-2 LL + 2p$
 - O termo $2p$ penaliza o modelo mais complexo
- Substituído o múltiplo de penalização por 3,841459
 - χ^2 com 1 grau de liberdade e 5% de significância
- P-valor = 0,05 considerado como valor crítico em cada iteração do stepwise

Dilema vício e variância



Validação cruzada k-fold

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avaliação
Iteração 1	VALIDAÇÃO	TREINO	TREINO	TREINO	TREINO	m_1
Iteração 2	TREINO	VALIDAÇÃO	TREINO	TREINO	TREINO	m_2
Iteração 3	TREINO	TREINO	VALIDAÇÃO	TREINO	TREINO	m_3
Iteração 4	TREINO	TREINO	TREINO	VALIDAÇÃO	TREINO	m_4
Iteração 5	TREINO	TREINO	TREINO	TREINO	VALIDAÇÃO	m_5



$$\frac{1}{5} \sum_{i=1}^5 m_i$$

Teste da razão da verossimilhança

- Modelo completo: 31 covariáveis
- Modelo restrito: 14 covariáveis
- $LRT < \chi^2$ com 17 graus de liberdade (restrições) e 5% de significância
- H_0 (17 estimativas estatisticamente = 0) não rejeitada
- A qualidade do ajuste não foi afetada com a retirada das covariáveis
- Optou-se por prosseguir com o modelo restrito

$$LRT = -2(LL_{\text{completo}} - LL_{\text{restrito}})$$



05

Análise de diagnóstico

Análise dos resíduos

- Resíduo: medida de afastamento de uma observação para o seu valor ajustado por um modelo
- Resíduos ordinários
- Verificação dos pressupostos através da normalidade dos resíduos
- Difíceis de generalizar para outras distribuições além da Normal

Resíduos quantílicos aleatorizados

- Se apresentam na forma da Normal Padrão, caso os parâmetros sejam estimados de forma consistente, independentemente da distribuição
- Baseado no teorema da inversa da função distribuição acumulada
- u_i é o valor da FDA do modelo proposto
- Φ^{-1} é a inversa da FDA da Normal Padrão
- Quando y é discreta, um recurso de aleatorização é aplicado

$$r_i = \Phi^{-1}(u_i)$$

Resíduos quantílicos aleatorizados

- Se $y_i = 0$, u_i assume um valor da distribuição uniforme entre 0 e $(1 - \hat{\pi}_i)$
- Se $y_i = 1$, u_i assume um valor da distribuição uniforme entre $(1 - \hat{\pi}_i)$ e 1
- Por fim, r_i assume o valor do quantil da Normal Padrão para $p = u_i$
- Dunn e Smyth (1996)

Gráfico dos resíduos x valores ajustados

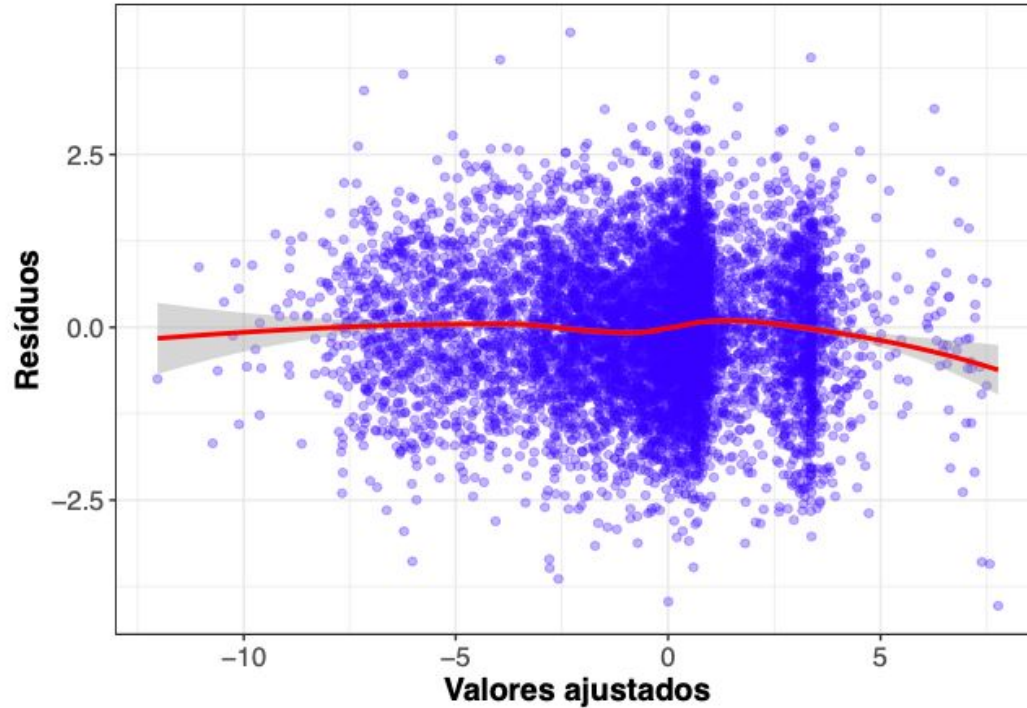
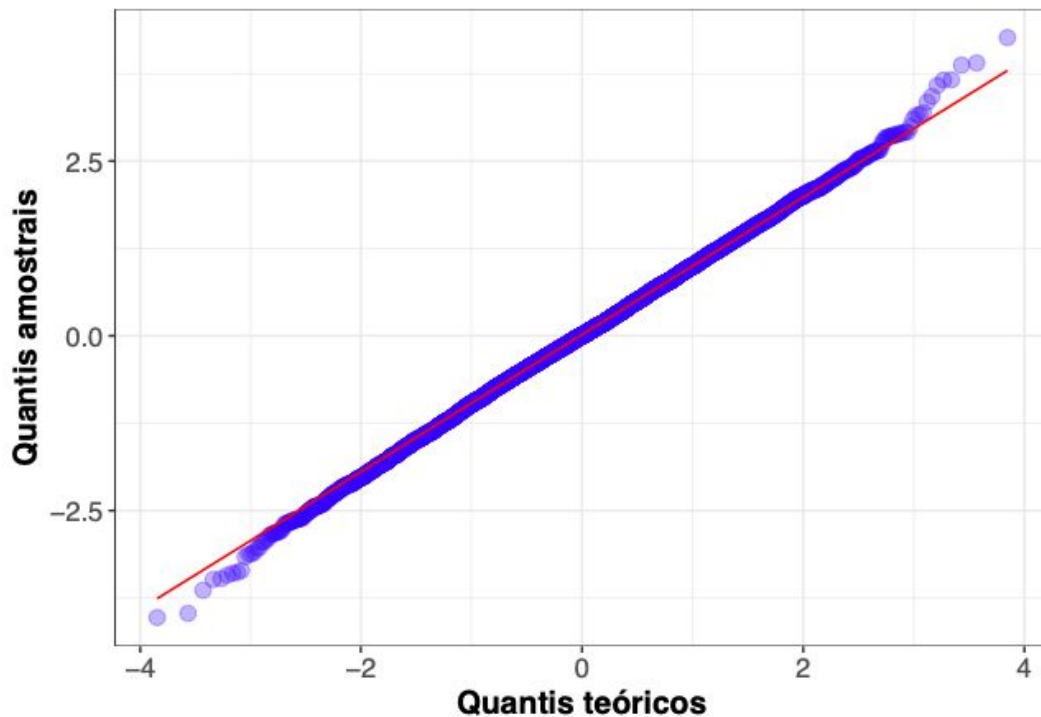


Gráfico quantil-quantil



Estimativas dos parâmetros

Tabela 4.: Estimativas dos parâmetros do modelo

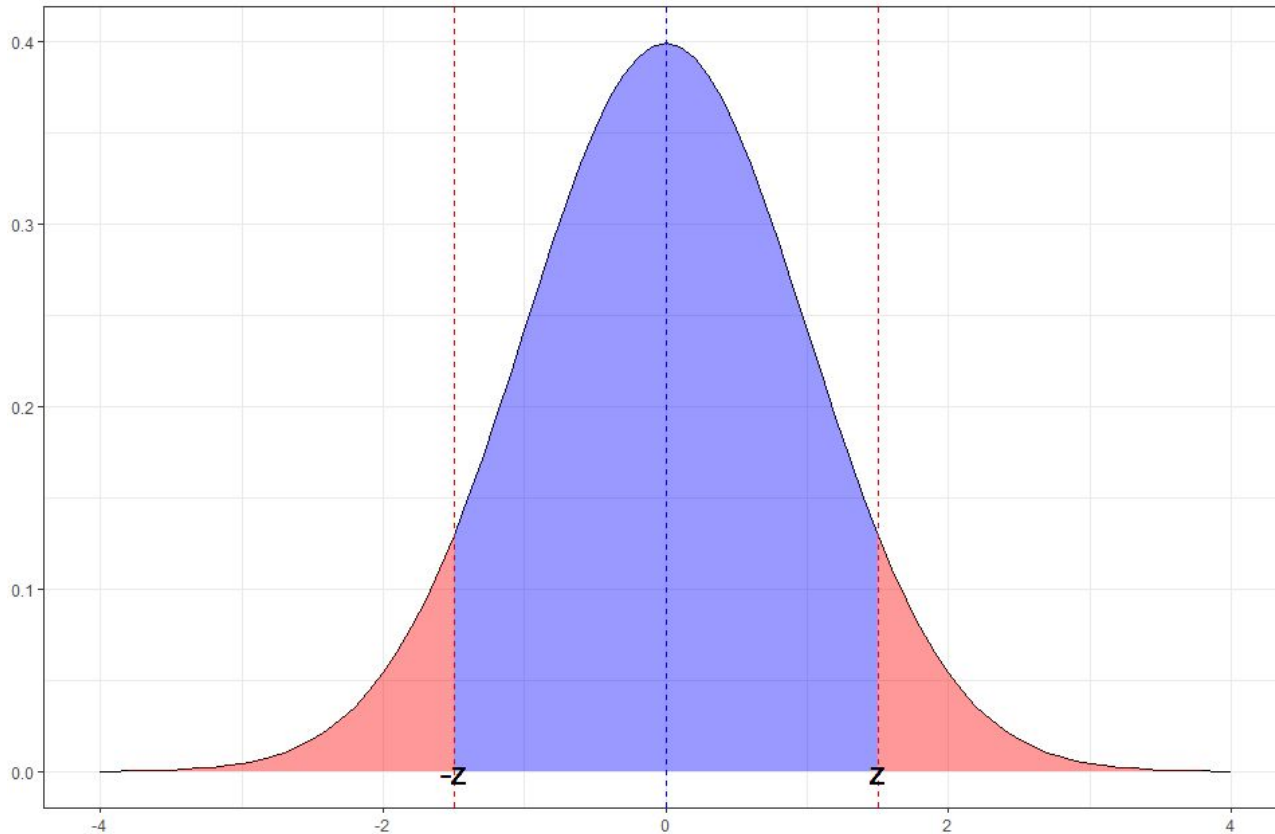
Covariável	Estimativa	Erro padrão	Wald	P-valor
Intercepto	1.5513	0.1316	11.7902	0.0000
X2	0.3693	0.0387	9.5437	0.0000
X4	-0.3558	0.0444	-8.0085	0.0000
X5	-0.2567	0.0487	-5.2749	0.0000
X6	-0.8296	0.1052	-7.8829	0.0000
X9	0.4532	0.1006	4.5028	0.0000
X11	-0.3950	0.0697	-5.6703	0.0000
X13	-2.6981	0.1237	-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

Significância das estimativas

- Teste Z de Wald
- Desvios padrões da estimativa com relação ao zero da normal padrão
- Verificar se há rejeição ou não de H_0
- O p-valor indica a probabilidade do parâmetro ser tão ou mais extremo que $|z|$

$$z_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\text{ep}(\hat{\beta}_j)} \quad H_0: \hat{\beta}_j = 0$$

Significância das estimativas



Razão de chances (odds ratio)

- Quanto se altera em média a chance de ocorrência do evento ao aumentar k unidades em determinada covariável?

Razão de chances (odds ratio)

- Quanto se altera em média a chance de ocorrência do evento ao aumentar k unidades em determinada covariável?

$$\ln(\text{odds}\{x\}) = \beta_0 + \beta_1 x$$

Razão de chances (odds ratio)

- Quanto se altera em média a chance de ocorrência do evento ao aumentar k unidades em determinada covariável?

$$\ln(\text{odds}\{x\}) = \beta_0 + \beta_1 x$$

$$\text{odds}\{x\} = e^{\beta_0 + \beta_1 x}$$

Razão de chances (odds ratio)

- Quanto se altera em média a chance de ocorrência do evento ao aumentar k unidades em determinada covariável?

$$\ln(\text{odds}\{x\}) = \beta_0 + \beta_1 x$$

$$\text{odds}\{x\} = e^{\beta_0 + \beta_1 x}$$

$$\text{OR}\{x + k, x\} = \frac{\text{odds}\{x + k\}}{\text{odds}\{x\}}$$

Razão de chances (odds ratio)

- Quanto se altera em média a chance de ocorrência do evento ao aumentar k unidades em determinada covariável?

$$\ln(\text{odds}\{x\}) = \beta_0 + \beta_1 x$$

$$\text{odds}\{x\} = e^{\beta_0 + \beta_1 x}$$

$$\text{OR}\{x + k, x\} = \frac{\text{odds}\{x + k\}}{\text{odds}\{x\}} = \frac{e^{\beta_0 + \beta_1 (x+k)}}{e^{\beta_0 + \beta_1 x}}$$

Razão de chances (odds ratio)

$$\text{OR}\{x+k, x\} = \frac{e^{\beta_0 + \beta_1(x+k)}}{e^{\beta_0 + \beta_1 x}}$$

Razão de chances (odds ratio)

$$\begin{aligned}\text{OR}\{x+k, x\} &= \frac{e^{\beta_0 + \beta_1(x+k)}}{e^{\beta_0 + \beta_1 x}} \\ &= e^{\beta_0 + \beta_1(x+k) - \beta_0 - \beta_1 x}\end{aligned}$$

Razão de chances (odds ratio)

$$\begin{aligned}\text{OR}\{x+k, x\} &= \frac{e^{\beta_0 + \beta_1(x+k)}}{e^{\beta_0 + \beta_1 x}} \\ &= e^{\beta_0 + \beta_1(x+k) - \beta_0 - \beta_1 x} \\ &= e^{\beta_1(x+k-x)}\end{aligned}$$

Razão de chances (odds ratio)

$$\begin{aligned}\text{OR}\{x+k, x\} &= \frac{e^{\beta_0 + \beta_1(x+k)}}{e^{\beta_0 + \beta_1 x}} \\ &= e^{\beta_0 + \beta_1(x+k) - \beta_0 - \beta_1 x} \\ &= e^{\beta_1(x+k-x)} \\ &= e^{k\beta_1}\end{aligned}$$

Razão de chances (odds ratio)

- Portanto, o aumento de k unidades em uma covariável x_j , *fixadas as demais*, multiplica a chance de ocorrência do evento de interesse por $e^{k\beta_j}$

$$\beta_j > 0 \implies e^{k\beta_j} > 1 \implies \pi_{x+k} > \pi_x$$

$$\beta_j < 0 \implies e^{k\beta_j} < 1 \implies \pi_{x+k} < \pi_x$$

Razão de chances (odds ratio)

- Considerando uma covariável categórica com níveis A, B e C, sendo A a categoria de referência:
- $OR\{B, A\} = odds\{B\} / odds\{A\}$
- $OR\{C, A\} = odds\{C\} / odds\{A\}$
- Portanto, estar na categoria B multiplica a chance de ocorrência do evento de interesse por e^{β_B} , com relação a estar na categoria A
- E estar na categoria C multiplica a chance de ocorrência do evento de interesse por e^{β_C} , com relação a estar na categoria A

Razão de chances (odds ratio)

- Ainda considerando a covariável categórica do exemplo anterior:
- $OR\{B, C\} = OR\{B, A\} / OR\{C, A\}$
- $OR\{C, B\} = OR\{C, A\} / OR\{B, A\}$
- Portanto, estar na categoria B multiplica a chance de ocorrência do evento de interesse por $e^{\beta_B - \beta_C}$, com relação a estar na categoria C
- E estar na categoria C multiplica a chance de ocorrência do evento de interesse por $e^{\beta_C - \beta_B}$, com relação a estar na categoria B

Interpretação das estimativas

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

Tabela 2.: Interpretação das métricas de desempenho

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

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

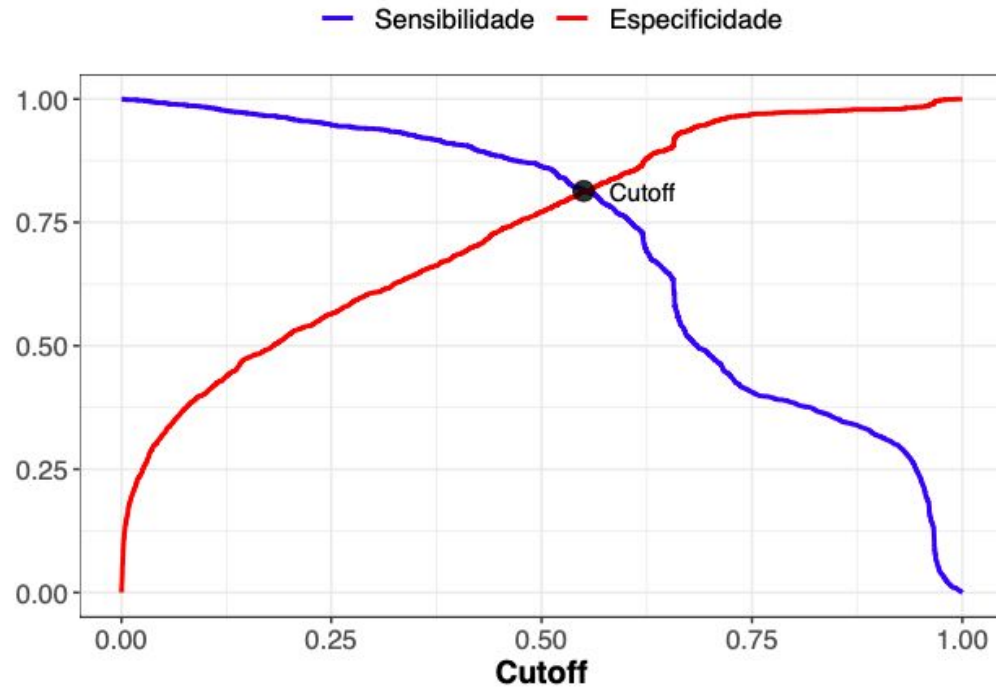
Covariável	Estimativa	Chance (0,5)	Variação (%)
X6	-0.8296	0.6605	-34
X9	0.4532	1.2543	25
X2	0.3693	1.2028	20
X4	-0.3558	0.8370	-16
X5	-0.2567	0.8795	-12



06

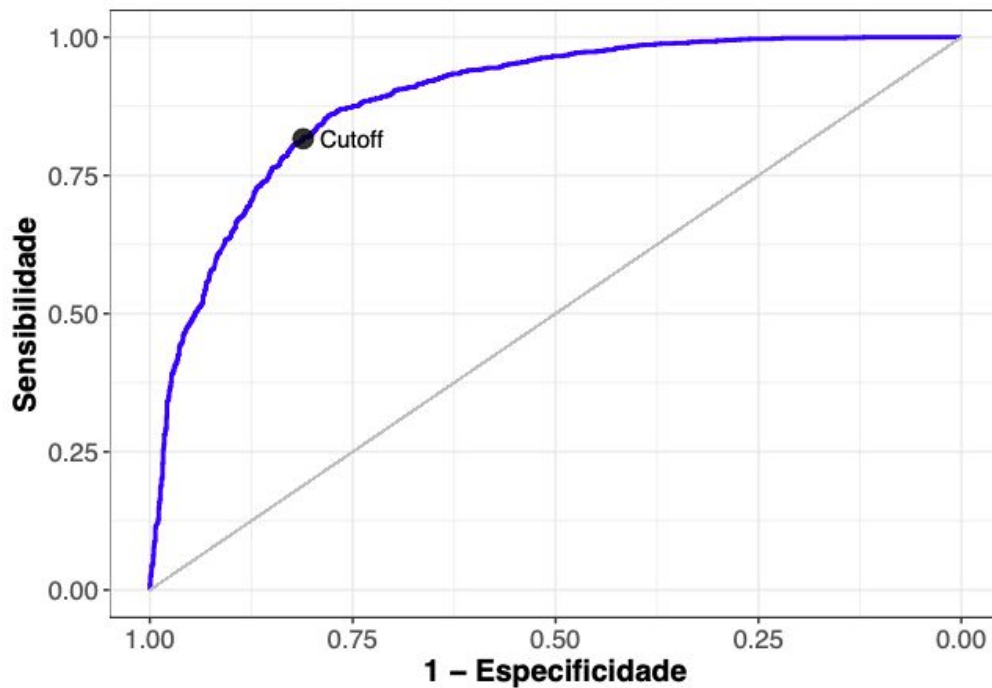
Avaliação do modelo

Escolha do cutoff



Curva ROC

AUC = 0,88



Matriz de confusão

Tabela 7.: Matriz de confusão

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

$$S = \frac{VP}{VP + FN}$$

$$E = \frac{VN}{VN + FP}$$

$$A = \frac{VN + VP}{VN + VP + FN + FP}$$

Matriz de confusão



Métricas de avaliação

- Acurácia:
 - Taxa de acerto geral
 - 82,79%
- Sensibilidade:
 - $P(\hat{y} = 1 \mid y = 1)$
 - 81,64%
- Especificidade:
 - $P(\hat{y} = 0 \mid y = 0)$
 - 81,36%



07

Considerações finais

Pontos importantes

- Conhecer / estudar o problema de negócio
- Explorar e garantir a qualidade dos dados (garbage in, garbage out)
- Escolher a técnica adequada para atingir os objetivos do negócio
- Considerar o “princípio da parcimônia”
- Garantir a qualidade do ajuste
- Atenção com o overfitting
- A saída do modelo faz sentido para o negócio?
- As métricas escolhidas fazem sentido para o negócio?
- Desenvolver uma estratégia de monitoramento do modelo

Recomendações



EstaTiDados

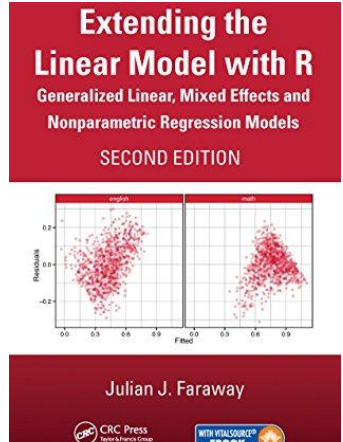
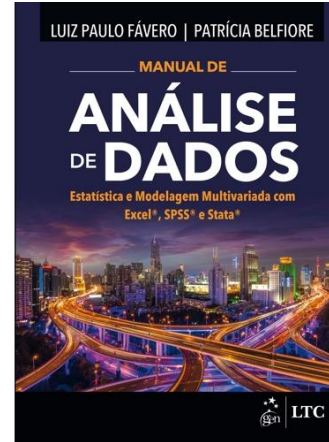
21,3 mil inscritos



StatQuest with Josh Starmer

390 mil inscritos

 Portal Action®



Universidade Federal do Paraná - Departamento de Estatística

CE225 - Modelos Lineares Generalizados


Prof. Cesar Augusto Taconeli

Recomendações

- [Estatidados](#)
- [StatQuest](#)
- [Portal Action](#)
- [Omega Data Science](#)
- [CE225 - Prof. Taconeli](#)
- [Manual de Análise de Dados](#)
- [Extending the Linear Model with R](#)

Artigo e códigos

- <https://github.com/juniorssz/dsbd-churn-analysis>



“Sem dados você é apenas
mais uma pessoa com uma
opinião”

—W. Edwards Demming

Obrigado

Contato:

<https://acsiunior.com/>

CREDITOS: Este template foi criado por **Slidesgo**.