Introdução ao Machine Learning

Dataprep e Classificação

May de 2022

Dataprep Parte I

Conteúdo

- Preditores categóricos
- Transformações 1:1
- Transformações 1:n
- Regressão Logística
- Matriz de Confusão
- Métricas de Classificação
- Curva ROC
- Múltiplas Notas de Corte

Preditores Categóricos

Preditor com apenas 2 categorias

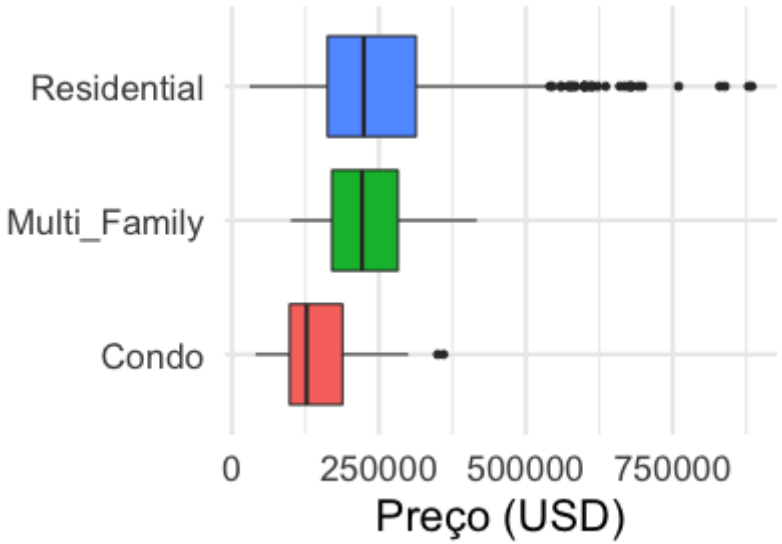
Saldo médio no cartão de crédito é diferente entre homens e mulheres?

$$y_i = \beta_0 + \beta_1 x_i \qquad \text{em que} \qquad x_i = \begin{cases} 1 & \text{se o i-ésimo carro for } \texttt{manual} \\ 0 & \text{se o i-ésimo carro for } \texttt{automático} \end{cases}$$

Ver ISL página 84 (Predictors with Only Two Levels).

Preditores Categóricos

Preditor com 3 ou mais categorias



Exemplo: Modelo linear

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Em que

$$x_{1i} = \begin{cases} 1 & \text{se for } \texttt{Multi\_Family} \\ 0 & \text{caso contrário} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{se for } \texttt{Residential} \\ 0 & \text{caso contrário} \end{cases}$$

Preditores Categóricos

Preditor com 3 ou mais categorias

"One hot enconding" ou "Dummies" ou "Indicadores".

type	(Intercept)	typeMulti_Family	typeResidential
Residential	1	0	1
Multi_Family	1	1	0
Condo	1	0	0
Residential	1	0	1
Condo	1	0	0
Multi_Family	1	1	0

steps: `step_dummy()`

Preditores Categóricos

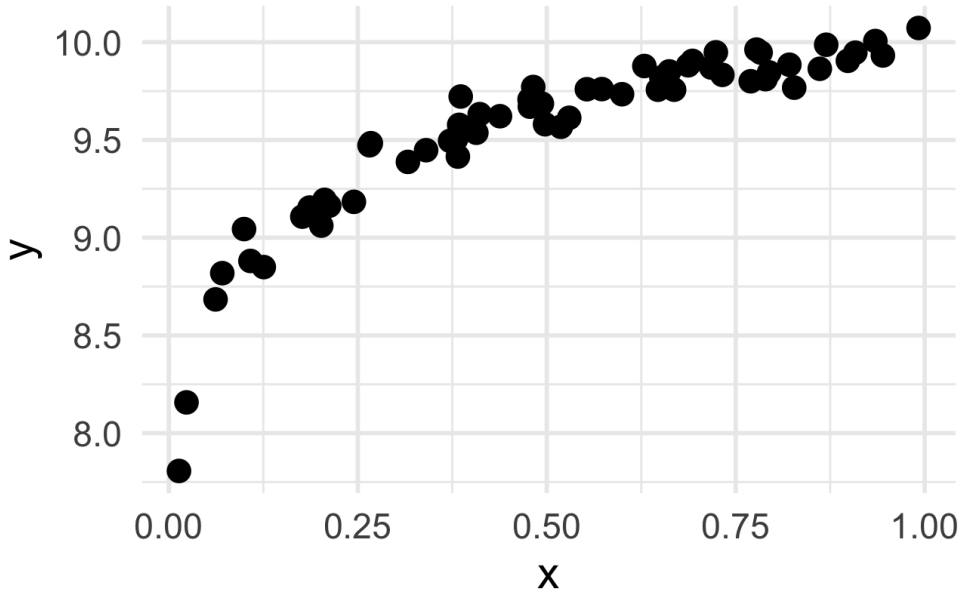
Preditor com 3 ou mais categorias

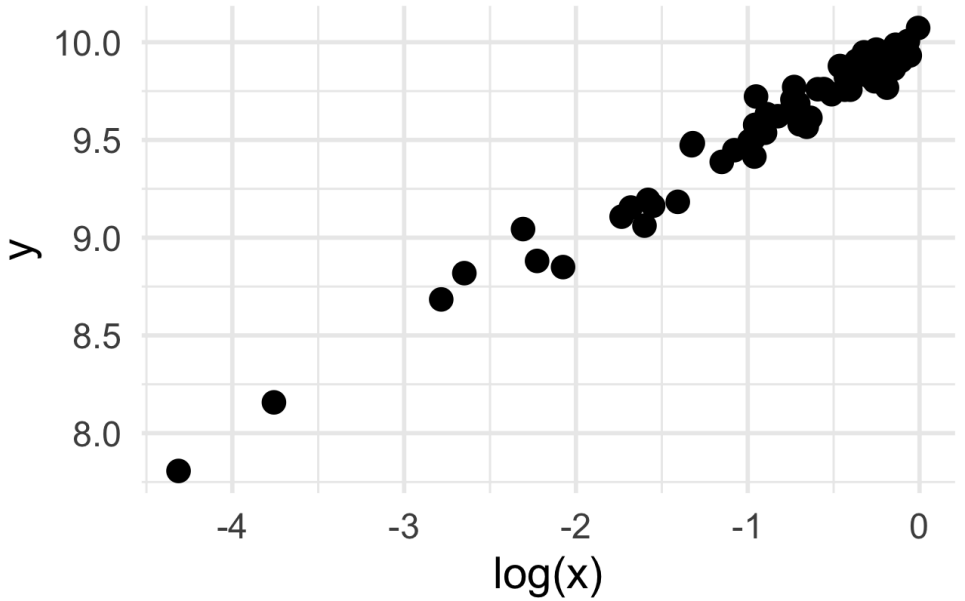
As previsões para cada categoria ficaria assim:

$$y_i = \begin{cases} \beta_0 & \text{se for } \texttt{Condo} \\ \beta_0 + \beta_1 & \text{se for } \texttt{Multi\_Family} \\ \beta_0 + \beta_2 & \text{se for } \texttt{Residential} \end{cases}$$

Transformações Não Lineares dos Preditores

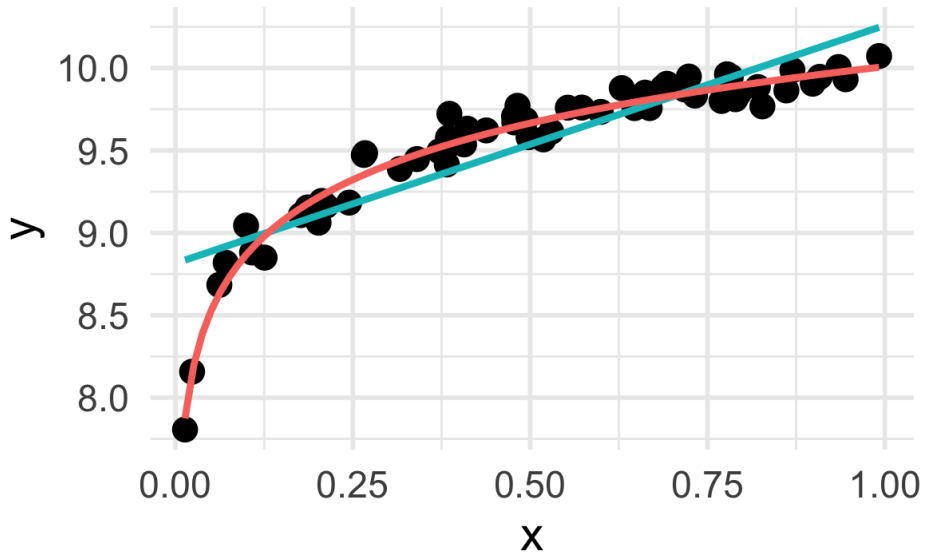
Exemplo: log





Transformações Não Lineares dos Preditores

Exemplo: log



Relação real entre x e y:
$y = 10 + 0.5log(x)$

Modelos propostos:

1) `y ~ x`

2) `y ~ log(x)`

Outras transformações comuns: raíz quadrada, Box-Cox.

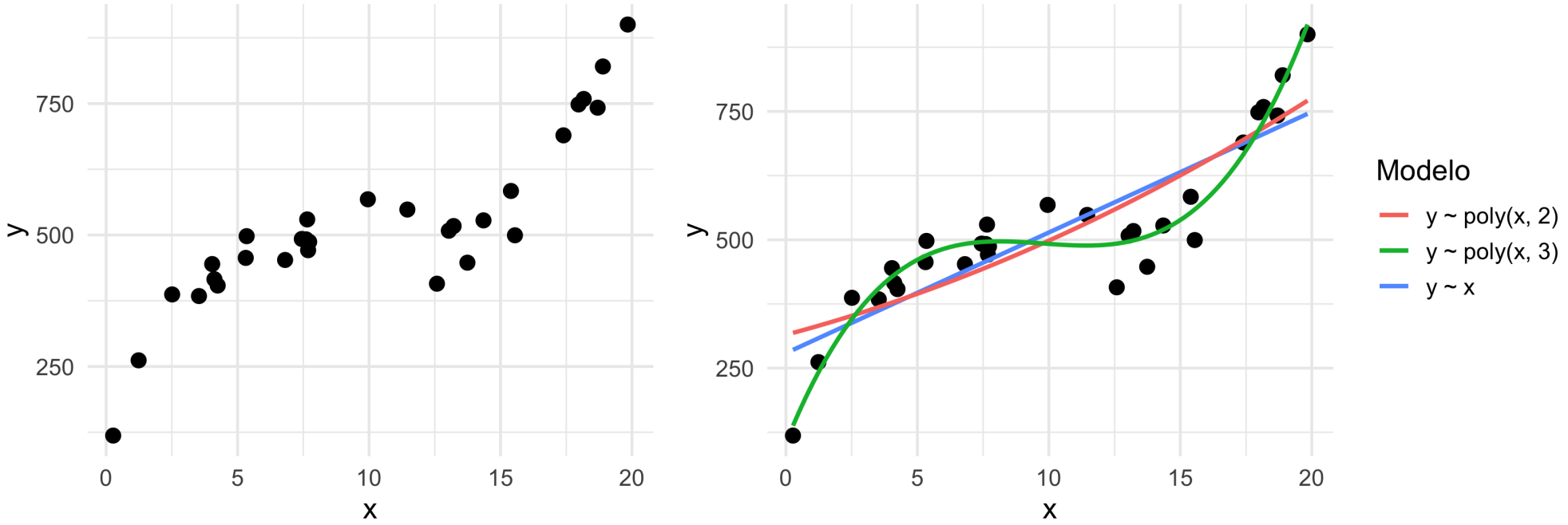
steps: `step_log()`, `step_BoxCox()`, `step_sqrt()`

Transformações Não Lineares dos Preditores

Exemplo: Regressão Polinomial

Relação real: $y = 500 + 0.4(x - 10)^3$

Modelo proposto:
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$



Transformações Não Lineares dos Preditores

Exemplo: Regressão Polinomial

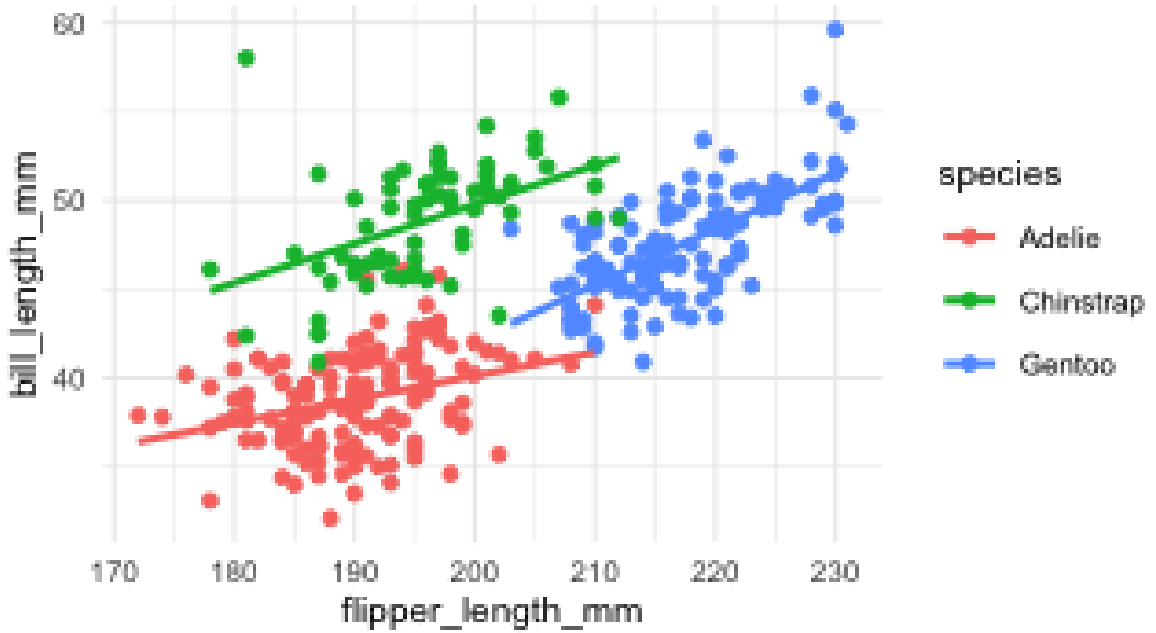
y	idade	idade2	idade3
456.5	5.3	28.2	149.7
492.5	7.4	55.4	412.2
548.4	11.5	131.3	1503.9
758.7	18.2	329.9	5993.0
444.7	4.0	16.3	65.6
748.3	18.0	322.8	5800.8
820.5	18.9	357.0	6744.3
517.0	13.2	174.7	2308.3

Outras expansões comuns: b-splines, natural splines.

steps: `step_poly()`, `step_bs()`, `step_ns`

Interações

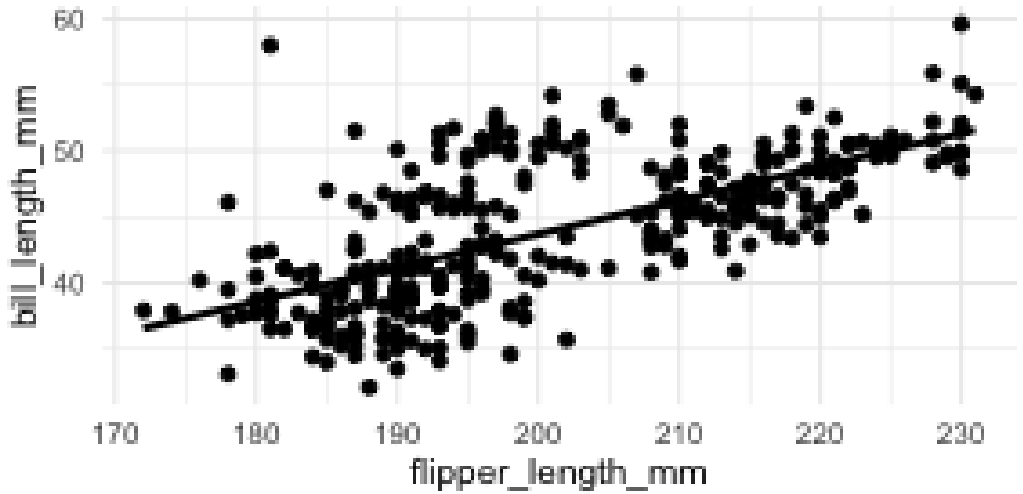
Interação entre duas variáveis explicativas: `species` e `bill_length_mm`



Interações

Modelo proposto (Matemático): Seja `y = flipper_length_mm` e `x = bill_length_mm`,

$$y = \beta_0 + \beta_1 x$$

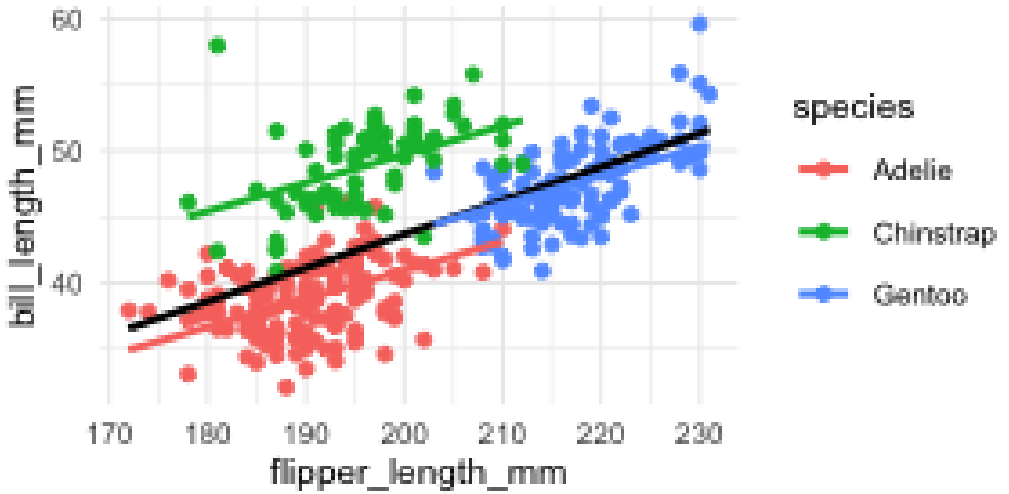


Modelo proposto (em R): `Sepal.Width ~ Sepal.Length`

Interações

Modelo proposto (Matemático): Seja `y = Sepal.Width` e `x = Sepal.Length`,

$$y = \beta_0 + \beta_1 x + \beta_2 I_{versicolor} + \beta_3 I_{virginica}$$

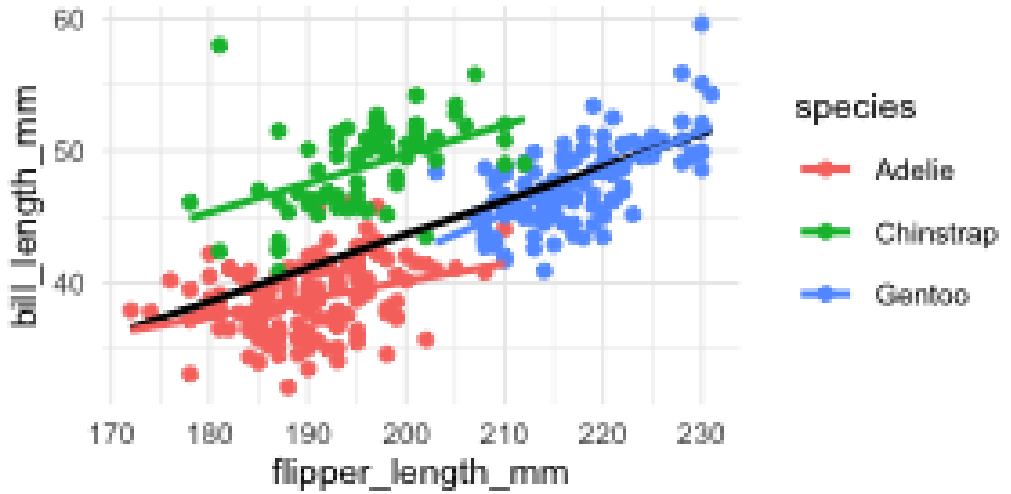


Modelo proposto (em R): `Sepal.Width ~ Sepal.Length + Species`

Interações

Modelo proposto (Matemático): Seja `y = Sepal.Width` e `x = Sepal.Length`,

$$y = \beta_0 + \beta_1 x + \beta_2 I_{versicolor} + \beta_3 I_{virginica} + \beta_4 x I_{versicolor} + \beta_5 x I_{virginica}$$



Modelo proposto (em R):
`step_interact(~flipper_length_mm:starts_with("species_"))`.

Exemplo 04

Outras referências

- Transformações recomendadas p/ cada modelo: https://www.tmwr.org/pre-proc-table.html
- Lista de transformações do recipes: https://recipes.tidymodels.org/reference/index.html
- Embbed: p/ quando o preditor tem muitas categorias: https://embed.tidymodels.org/
- Textos: quando colunas tem textos https://github.com/tidymodels/textrecipes
- Séries temporais: https://business-science.github.io/timetk/reference/index.html#section-feature-engineering-operations-recipe-steps-

Classificação

Regressão Logística

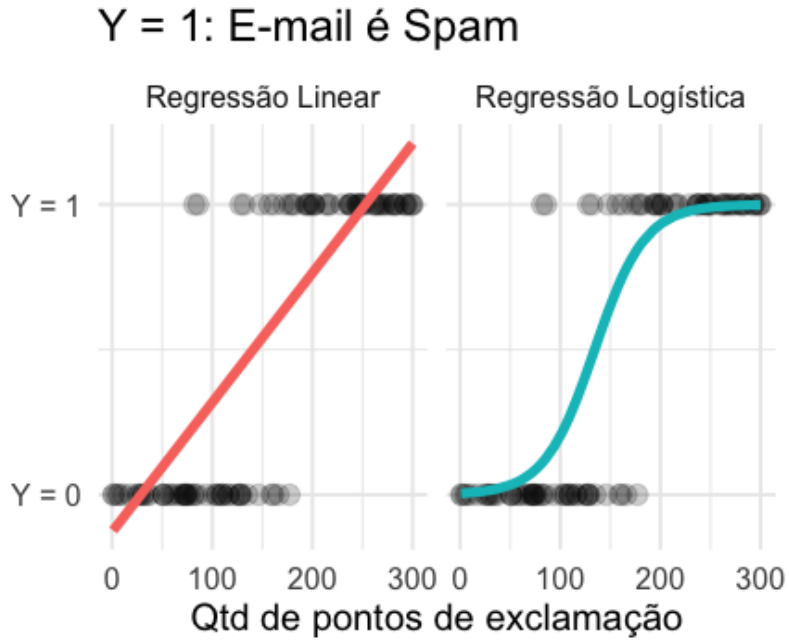
Para $Y \in \{0, 1\}$ (binário)

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Ou...

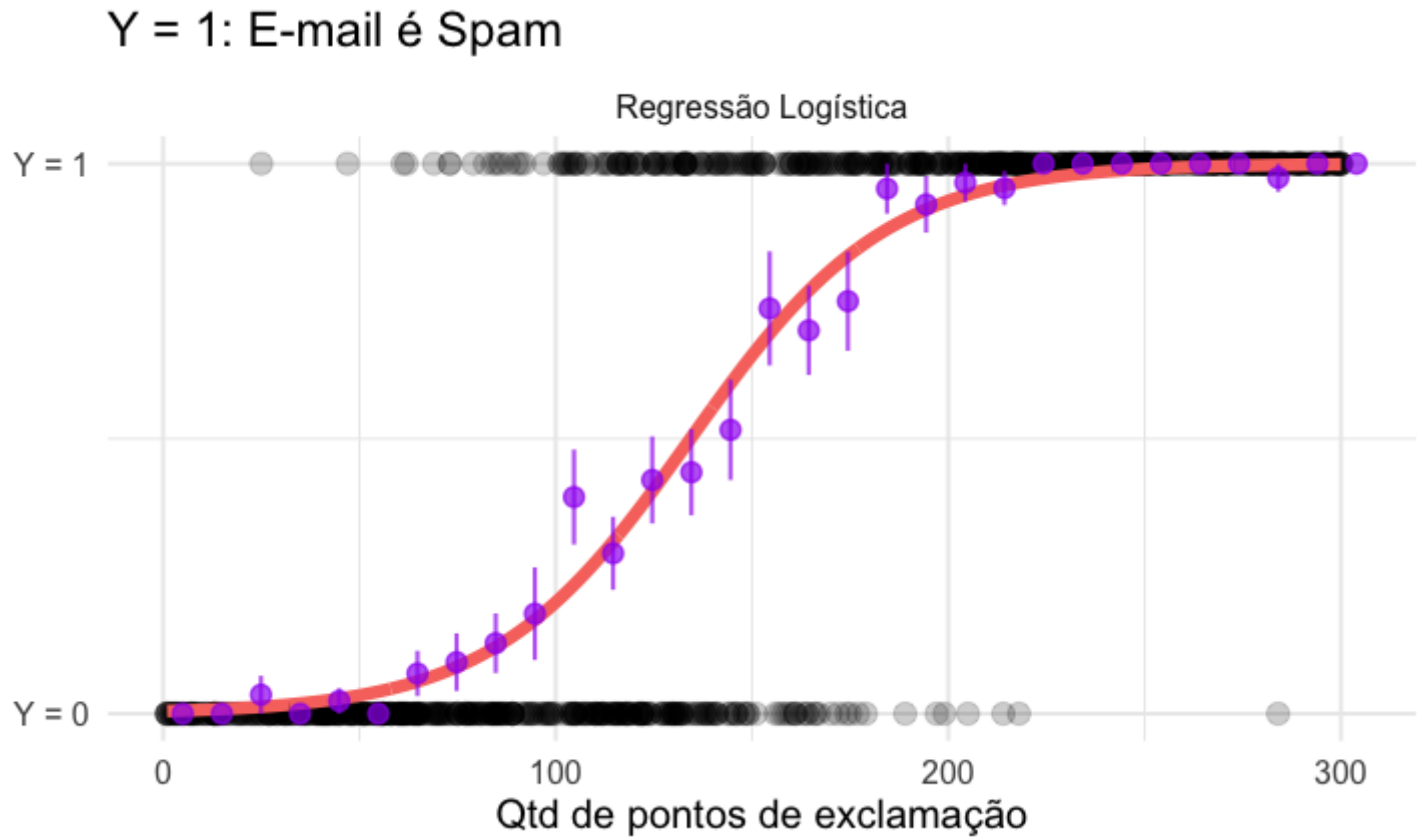
$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

```
### No R:
logistic_reg() %>%
  fit(spam ~ exclamacoes, data = dt_
```

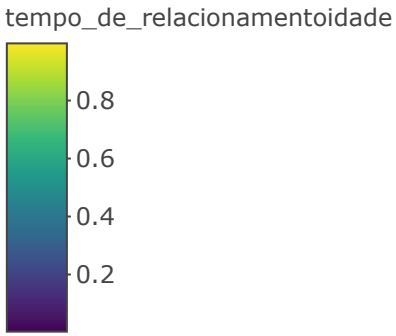


Ver ISL página 131 (Logistic Regression).

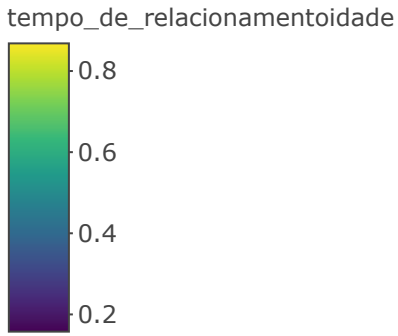
Regressão Logística



Regressão Logística



Árvore de Decisão



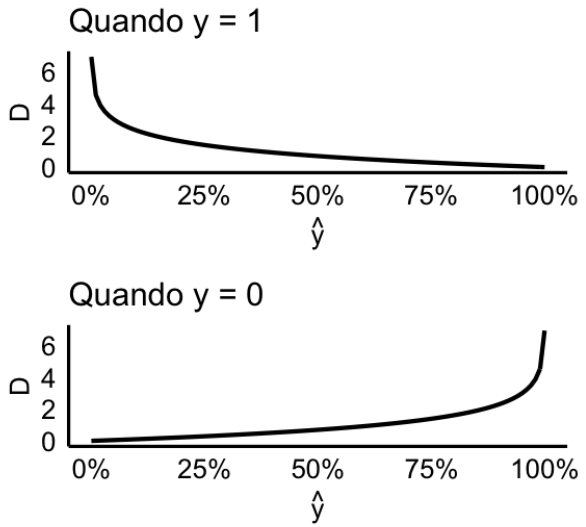
Regressão Logística - Custo

A **Métrica** que a regressão logística usa de **Função de Custo** chama-se *log-loss* (ou *Binary Cross-Entropy*):

$$D = \frac{-1}{N} \sum [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

Para cada linha da base de dados seria assim...

$$D_i = \begin{cases} -\log(\hat{y}_i) & \text{quando } y_i = 1 \\ -\log(1 - \hat{y}_i) & \text{quando } y_i = 0 \end{cases}$$



Regressão Logística - Regularização

A **Métrica** que a regressão logística usa de **Função de Custo** chama-se *log-loss* (ou *Binary Cross-Entropy*):

$$D = \frac{-1}{N} \sum [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

Regularizar é analogo a Regressão Linear.

$$D_{regularizado} = D + \lambda \sum_{j=1}^{p} |\beta_j|$$

PS1: Se $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 x$ então $\hat{p}_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$.

Regressão Logística - Predições

O "produto final" será um vetor de probabilidades estimadas.

pts excl	classe observada	prob	classe predita
167	Spam	**0.79**	**Spam**
129	Spam	**0.45**	**Não Spam**
299	Spam	**1.00**	**Spam**
270	Spam	**1.00**	**Spam**
187	Spam	**0.89**	**Spam**
85	Não Spam	**0.12**	**Não Spam**

Matriz de Confusão

	Observado	
Predito	**Neg**	**Pos**
Neg	TN	FN
Pos	FP	TP

p > 50%	Observado	
Predito	**Não Spam**	**Spam**
Não Spam	410	73
Spam	52	465

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{recall/TPR} = \frac{TP}{TP+FN}$$

$$\text{F1 score} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

$$\text{FPR} = \frac{FP}{FP+TN}$$

Nota de Corte (Threshold)

p > 10%	Observado	
Predito	**Não Spam**	**Spam**
Não Spam	267	8
Spam	195	530

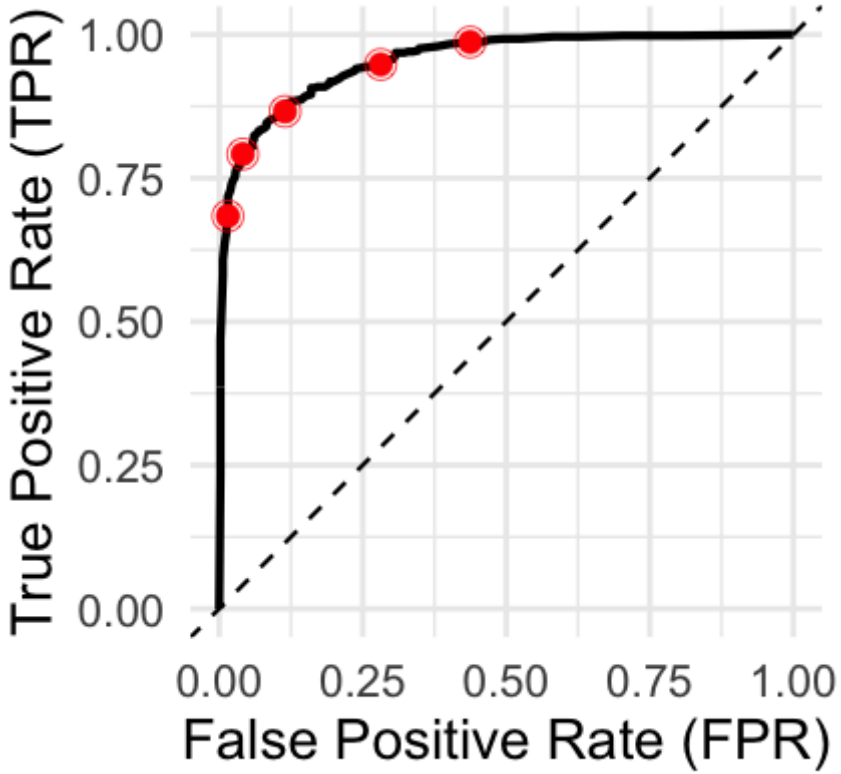
p > 25%	Observado	
Predito	**Não Spam**	**Spam**
Não Spam	332	28
Spam	130	510

p > 50%	Observado	
Predito	**Não Spam**	**Spam**
Não Spam	410	73
Spam	52	465

p > 75%	Observado	
Predito	**Não Spam**	**Spam**
Não Spam	443	112
Spam	19	426

p > 90%	Observado	
Predito	**Não Spam**	**Spam**
Não Spam	456	171
Spam	6	367

Curva ROC

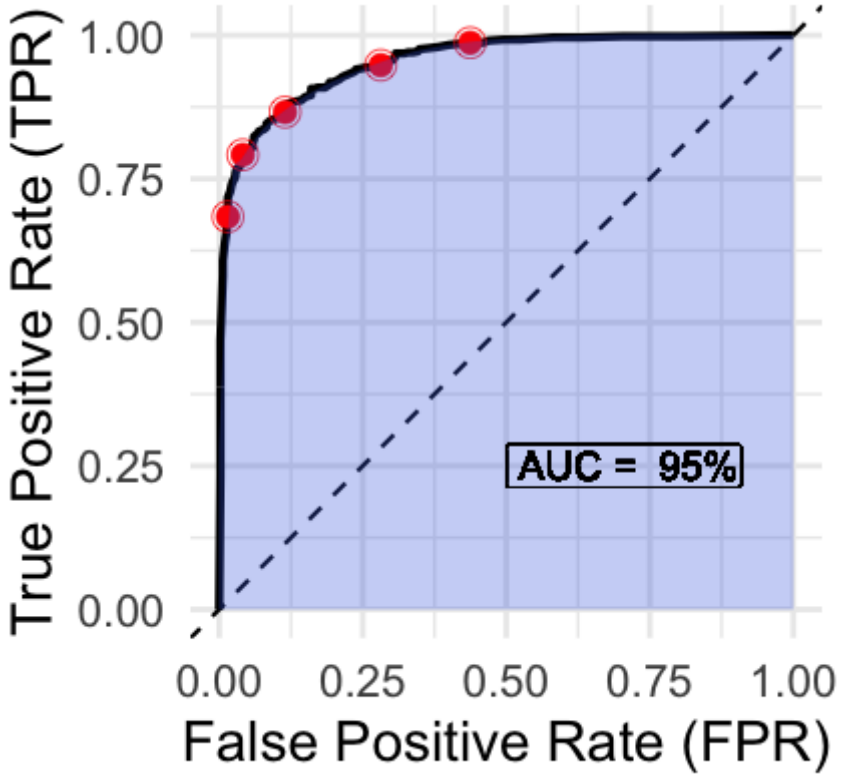


	Observado	
Predito	**Neg**	**Pos**
Neg	TN	FN
Pos	FP	TP

$$\text{TPR} = \frac{TP}{TP+FN}$$

$$\text{FPR} = \frac{FP}{FP+TN}$$

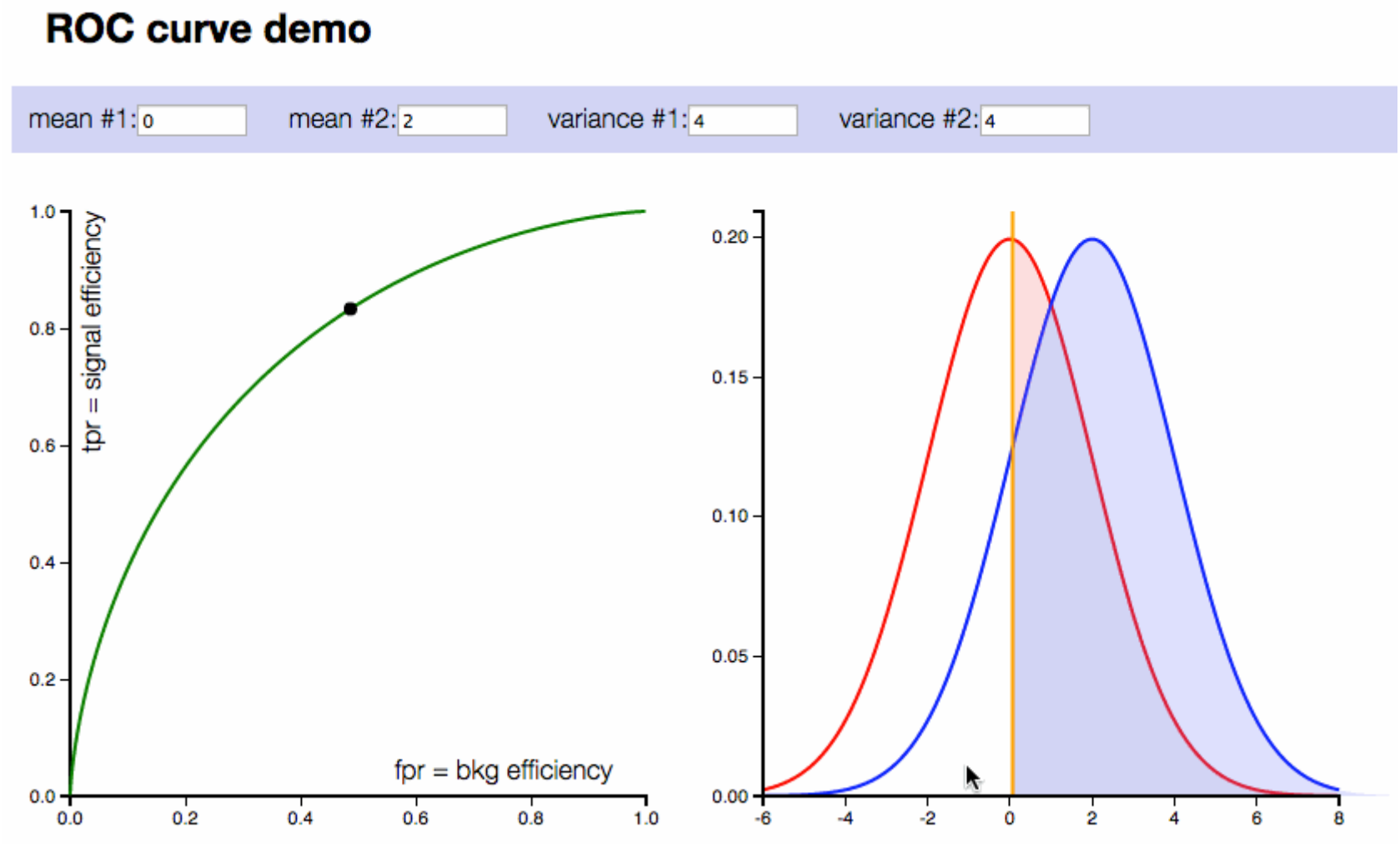
Curva ROC - Métrica AUC



	Observado	
Predito	**Neg**	**Pos**
Neg	TN	FN
Pos	FP	TP

$$\text{AUC} = \text{Area Under The ROC Curve}$$

Curva ROC - Playground



Múltiplas Notas de Corte

Risco por Segmentação

	Observado			
Predito	**Neg**	**Pos**	**N**	**Risco**
A (até 0,19)	90	11	101	11%
B (até 0,44)	60	40	100	40%
C (até 0,62)	39	60	99	60%
D (0,62 ou +)	20	80	100	80%

Usamos o `score` como preferirmos

```
dados %>%
  mutate(
    segmento = case_when(
      score <  0.19 ~ "A",
      score <  0.44 ~ "B",
      score <  0.62 ~ "C",
      score >= 0.62 ~ "D"))
```

Exemplo 05

Exercício 02