Aprendizagem estatística em altas dimensões

Florencia Leonardi

Conteúdo

- Árvores de regressão e classificação
- Agregação por bootstrap - *Bagging*
- Florestas aleatórias
- Modelos aditivos - *Boosting*

Árvores de decisão

- Este método devolve uma função $g : \mathcal{X} \to \mathcal{Y}$ obtida a partir dos dados $\mathcal{D}$,

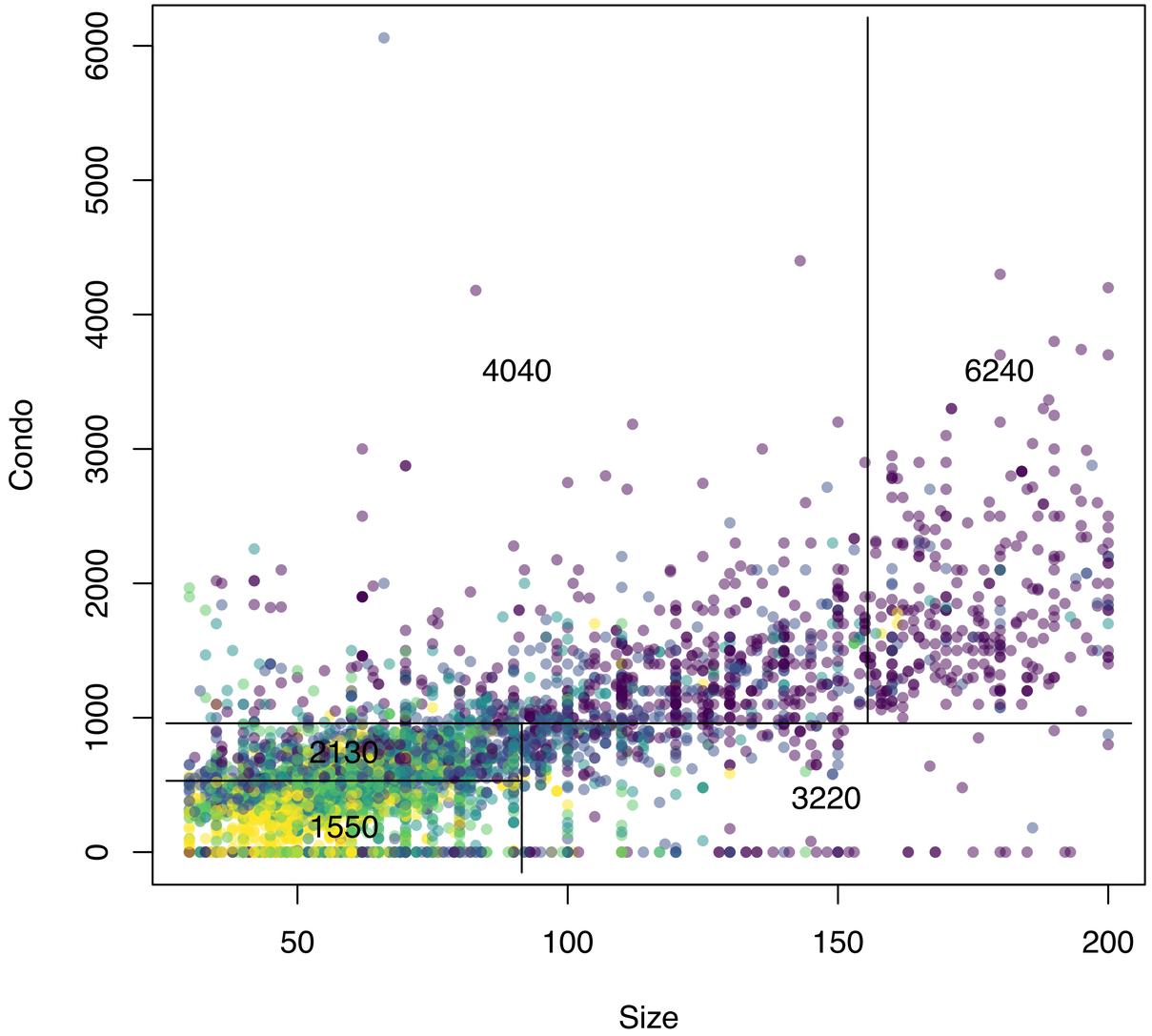
  mas que não está baseada em nenhum modelo paramétrico (é um método totalmente não paramétrico)

- A construção da função g está baseada em sucessivas divisões do espaço de variáveis preditoras em regiões simples (retângulos)

- Estas divisões sucessivas podem ser descritas graficamente por meio de uma árvore

- Estas árvores de decisão não costumam ter, sozinhas, uma grande acurácia nas predições, mas combinadas levam a métodos poderosos (florestas aleatórias, bagging, boosting...)

Árvores de regressão

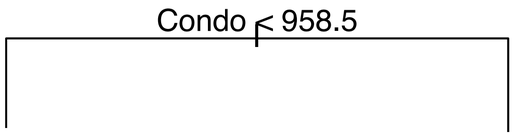


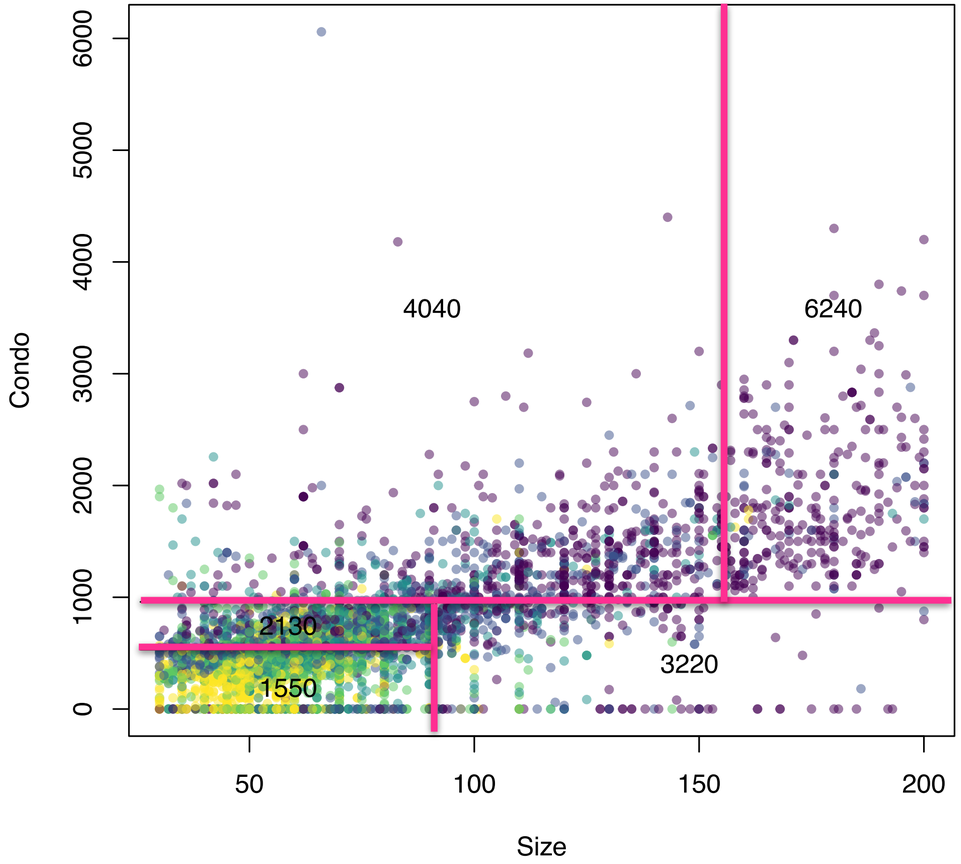
O objetivo inicial é encontrar regiões relativamente simples $R_1, \ldots, R_J$ no espaço $\mathcal{X}$ das variáveis preditoras que minimizem o erro:

$$\widehat{E}_D(R_1, \ldots, R_J) = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \bar{y}_{R_j})^2$$

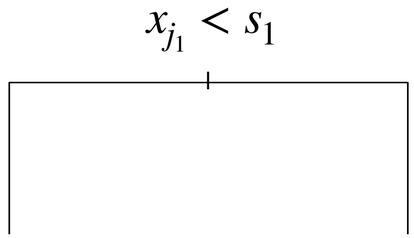
Como fazer isso de forma eficiente?

Árvores de regressão





Árvores de regressão



Para $j = 1, \ldots, p$ e $s \in \mathbb{R}$ definimos o par de semi-planos

$$R_1(j, s) = \{x \in \mathbb{R}^p : x_j < s\} \text{ e } R_2(j, s) = \{x \in \mathbb{R}^p : x_j \geq s\}$$

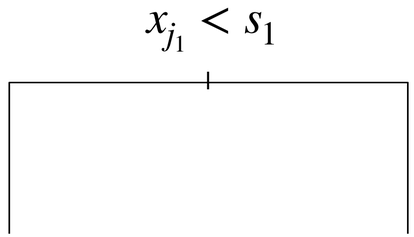
Procuramos o valor de j e s que minimizem o erro:

$$\sum_{i \in R_1(j,s)} (y_i - \bar{y}_{R_1(j,s)})^2 + \sum_{i \in R_2(j,s)} (y_i - \bar{y}_{R_2(j,s)})^2$$

Suponhamos que os valores obtidos foram $j = j_1$ e $s = s_1$.

Com eles fazemos a primeira divisão binária na árvore

Árvores de regressão



Repetimos o procedimento dentro de cada região obtida na divisão anterior do espaço.

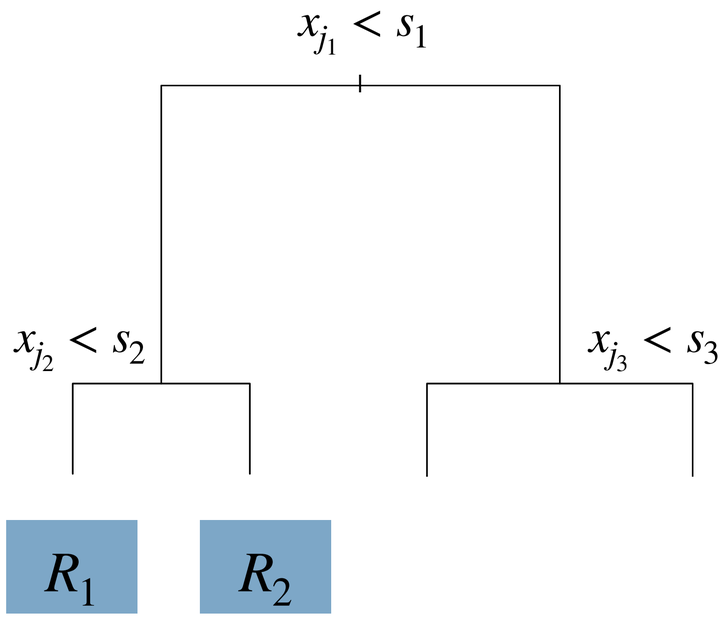
Ou seja escolhemos j_2 e s_2 que minimizem

$$\sum_{i \in R_1(j_1,s_1) \cap R_1(j_2,s_2)} (y_i - \bar{y}_{R_1(j_2,s_2)})^2 + \sum_{i \in R_1(j_1,s_1) \cap R_2(j_2,s_2)} (y_i - \bar{y}_{R_2(j_2,s_2)})^2$$

e j_3 e s_3 que minimizem

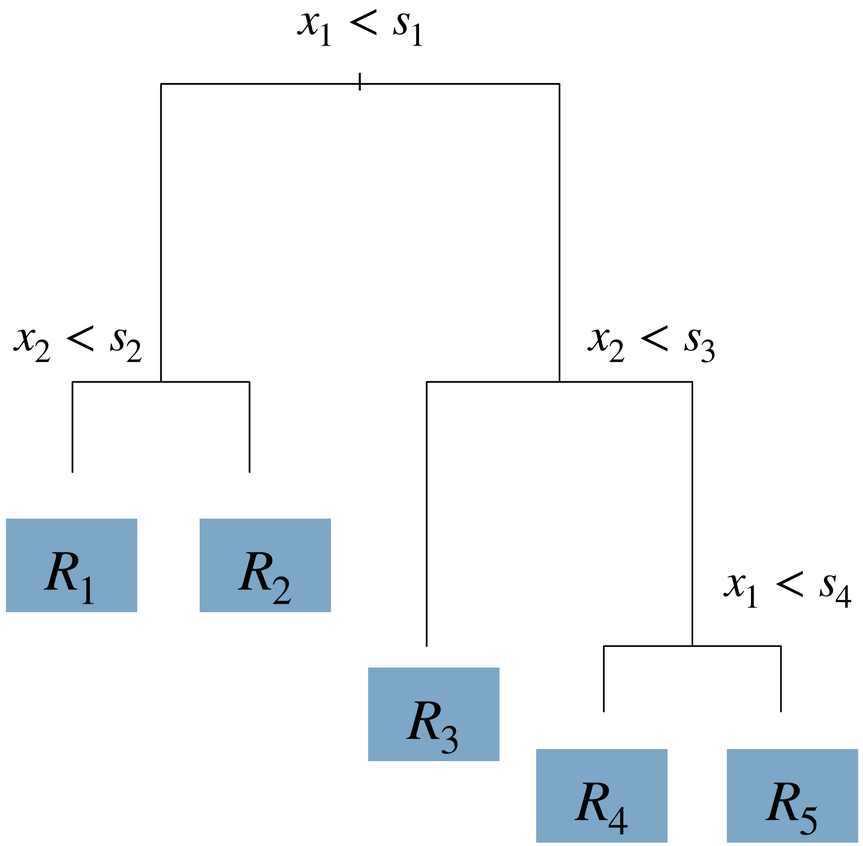
$$\sum_{i \in R_2(j_1,s_1) \cap R_1(j_3,s_3)} (y_i - \bar{y}_{R_1(j_3,s_3)})^2 + \sum_{i \in R_2(j_1,s_1) \cap R_2(j_3,s_3)} (y_i - \bar{y}_{R_2(j_3,s_3)})^2$$

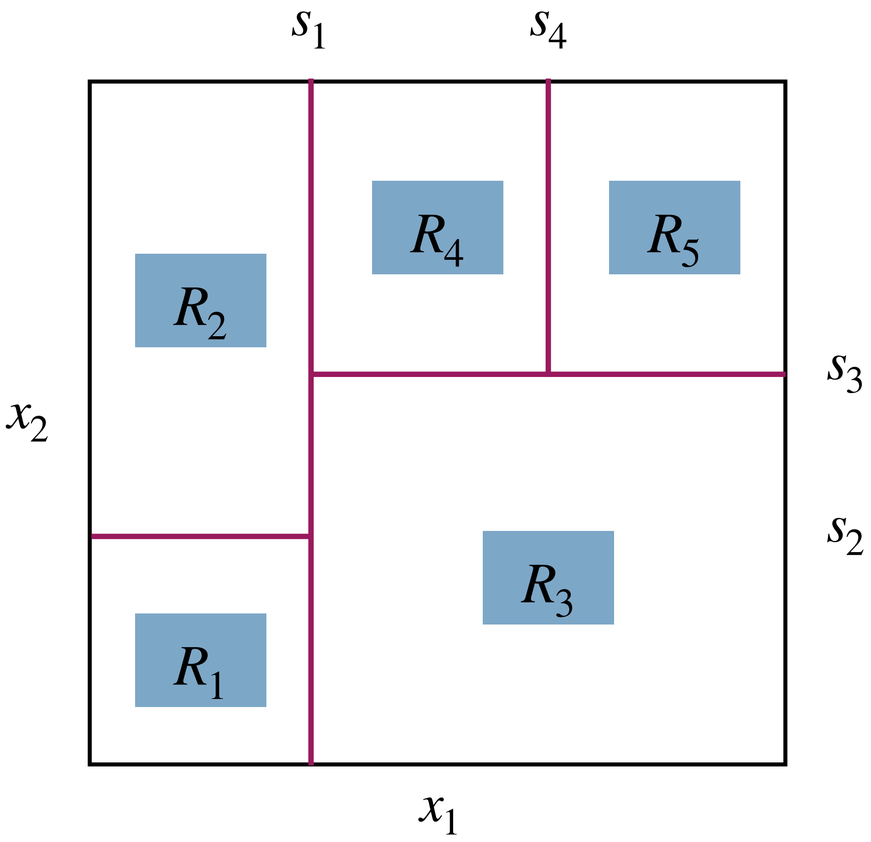
Árvores de regressão



O mesmo procedimento é iterado até que não encontramos mais regiões com um número mínimo de observações.

Árvores de regressão

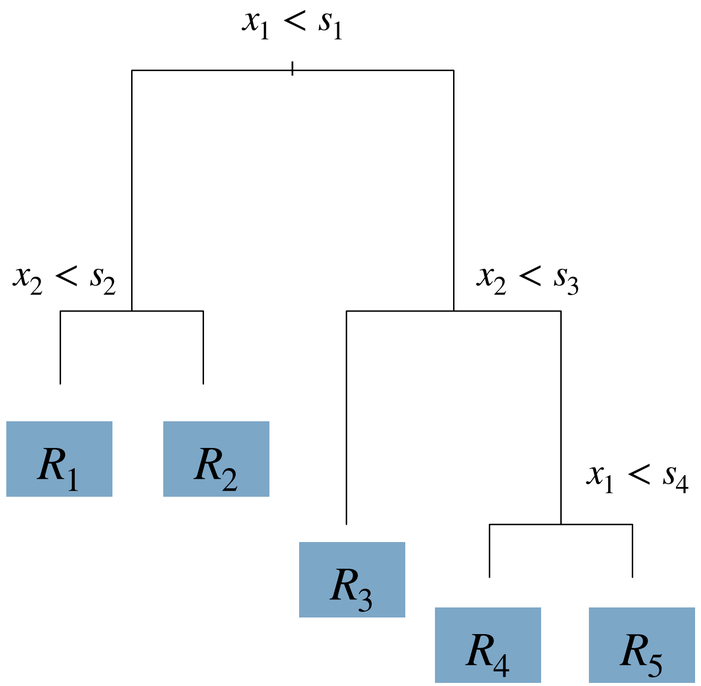


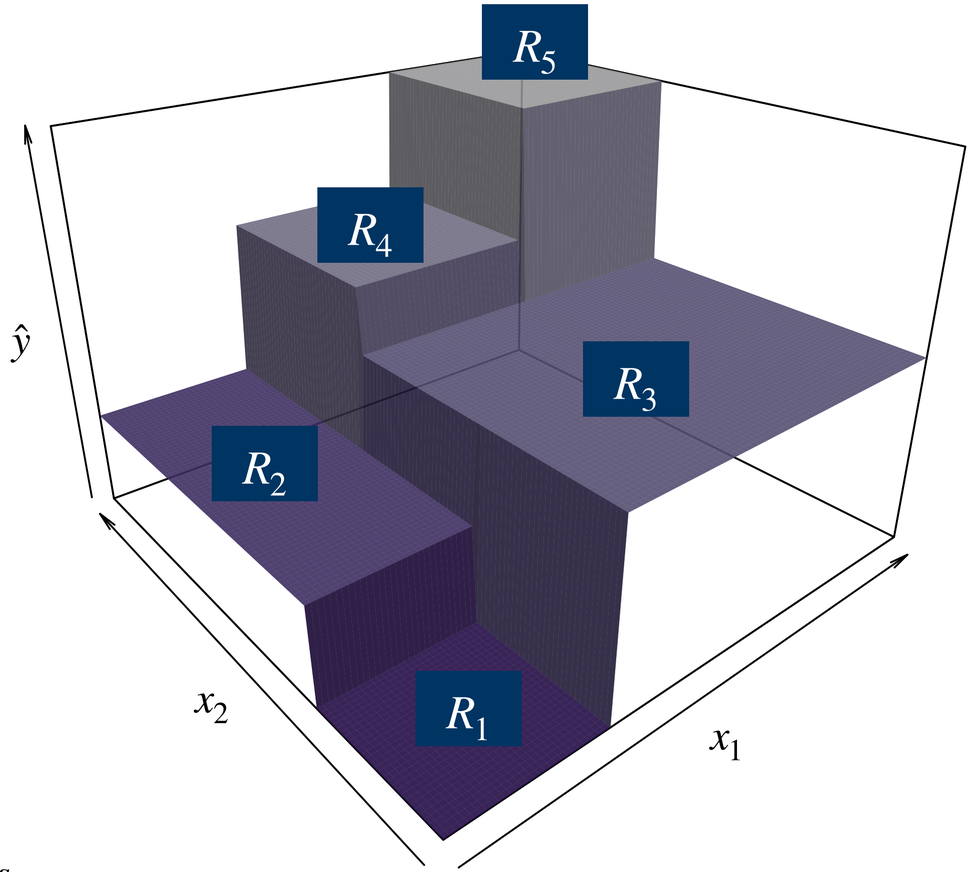


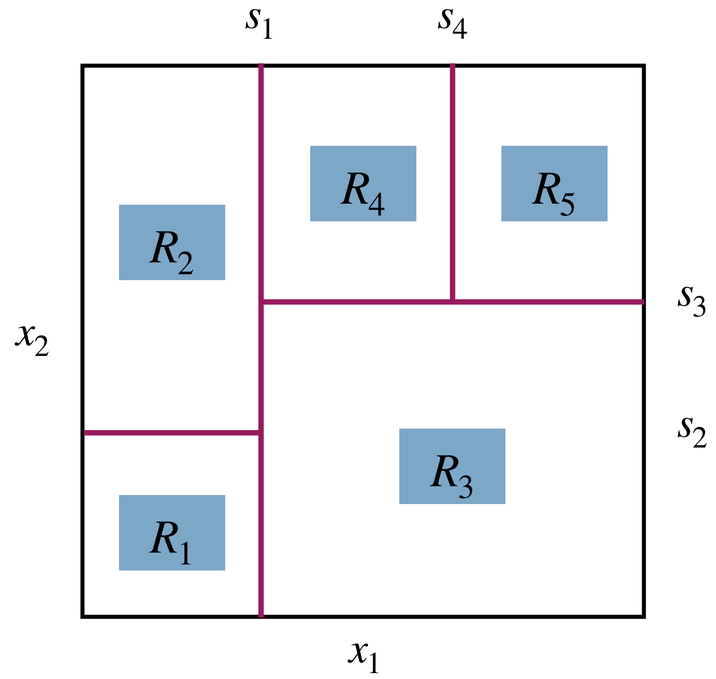
Árvores de regressão

$$T = (R_1, \ldots, R_5)$$

$$|T| = 5$$







Árvores de regressão

- O processo descrito anteriormente muito provavelmente vai superajustar a amostra, se em cada região houver poucas observações (o caso extremo disso seria ter regiões com uma única observação, em cujo caso o erro dentro da amostra seria 0!)
- Por outro lado, se fixarmos um número grande de observações por região, as regiões serão grandes e os pontos dentro de cada região estarão afastados, o que pode levar a regiões muito heterogêneas e a uma estimativa ruim da média dentro de cada região (aumentando o viés do modelo).
- Uma forma de evitar estes problemas é “podar” a árvore final para ter um certo balanço entre ajuste e complexidade

Poda da árvore

Como em muitas outras abordagens, a poda da árvore está baseada na regularização do erro estimado dentro da amostra

Para cada $\alpha > 0$ escolhemos a árvore T que minimiza

$$\widehat{E}_D(T) + \alpha |T| = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \bar{y}_{R_j})^2 + \alpha |T|$$

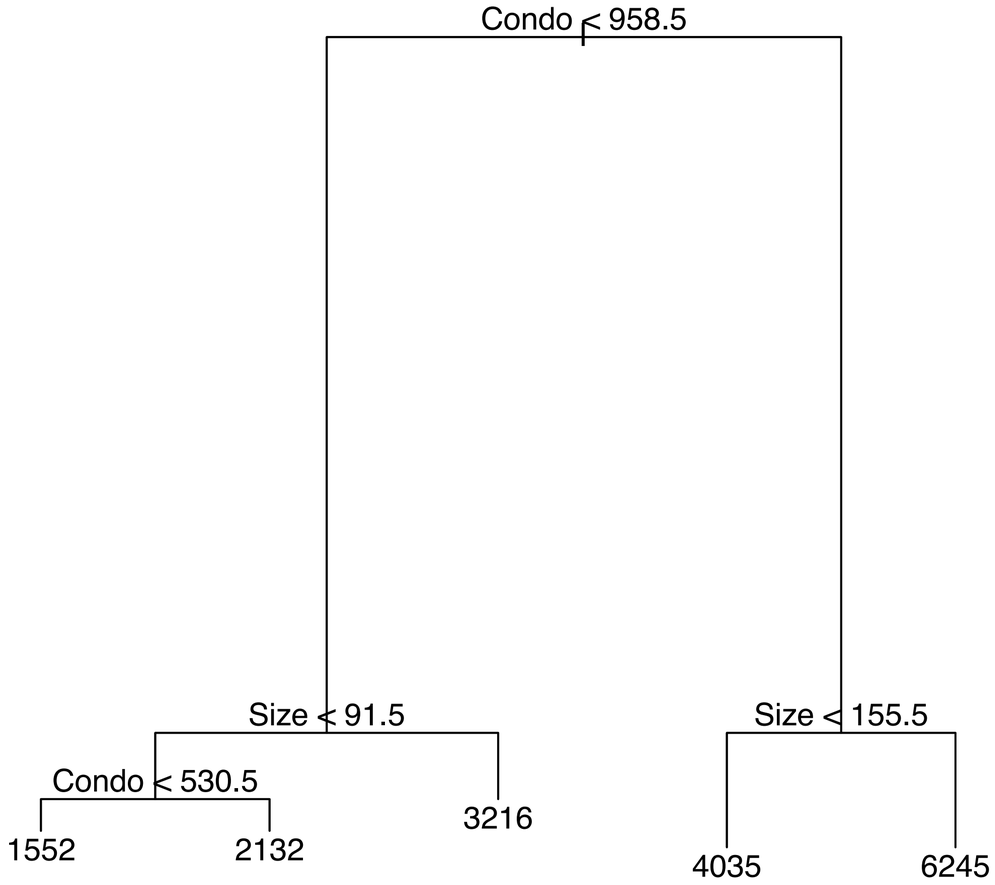
Quando α aumenta, mais ramos são “podados” da árvore e menos regiões são obtidas, diminuindo a variância e aumentando o viés do modelo.

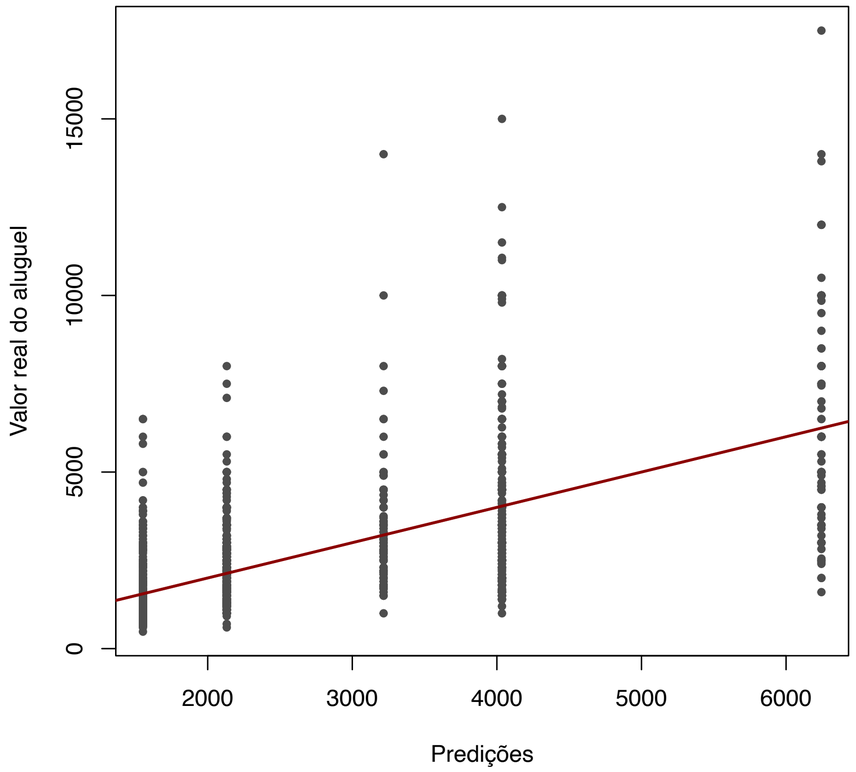
O valor ótimo de α pode ser escolhido com um conjunto de validação ou por validação cruzada

Algoritmo: árvore de regressão

1. Utilize divisão binária recursiva para construir uma árvore nos dados de treinamento, de tal forma que cada região obtida contenha um número mínimo de observações.
2. Pode a árvore obtida no passo anterior mudando o valor de α, de tal forma de obter uma sequência de árvores $T_1, \ldots, T_M$
3. Escolha uma das árvores $T_1, \ldots, T_M$ por validação cruzada

Árvores de regressão





$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 1543.43$$

Árvores de classificação

O objetivo é encontrar regiões $R_1, \ldots, R_J$ no espaço $\mathcal{X}$ das variáveis preditoras que minimizem o erro estimado:

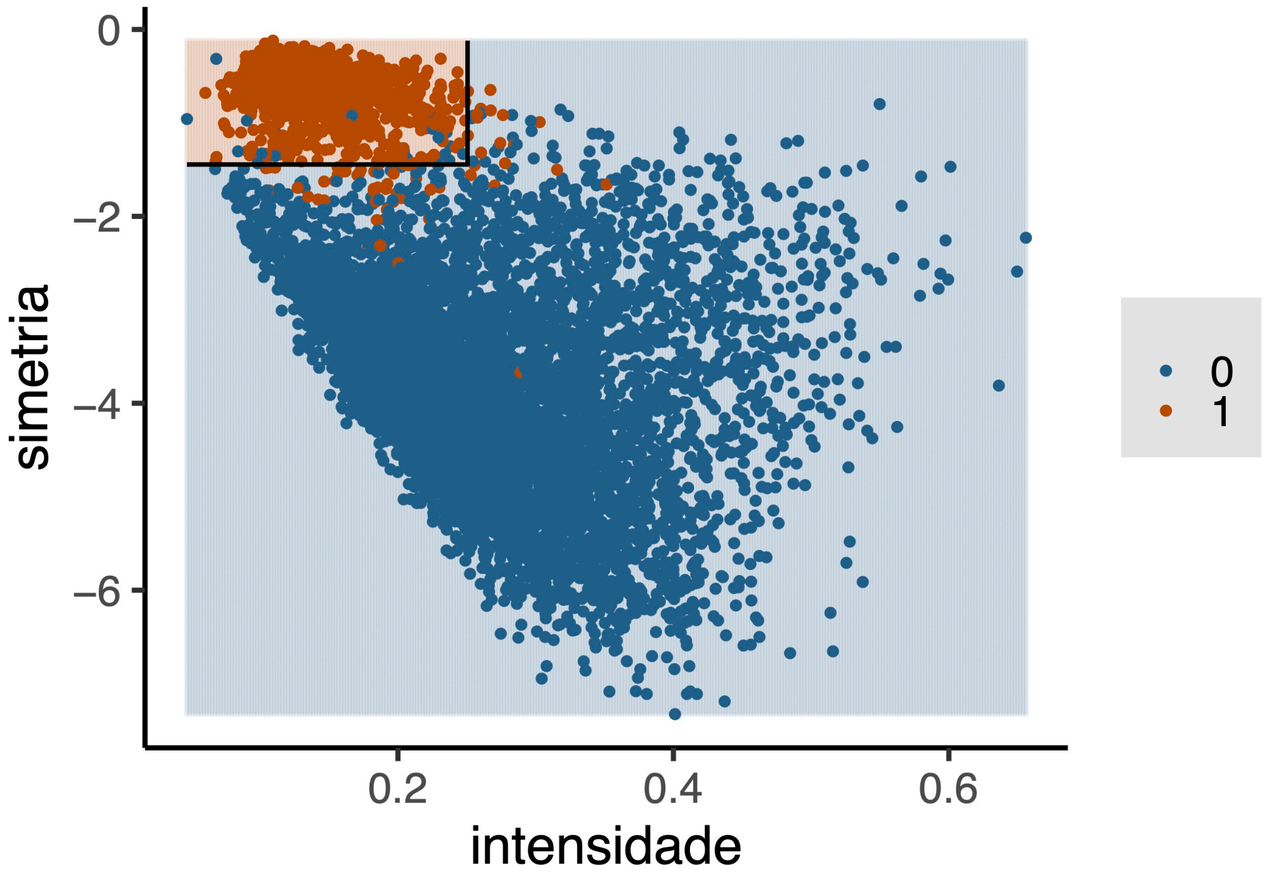
$$\widehat{E}_D(R_1, \ldots, R_J) = \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$$

→ *índice de Gini*

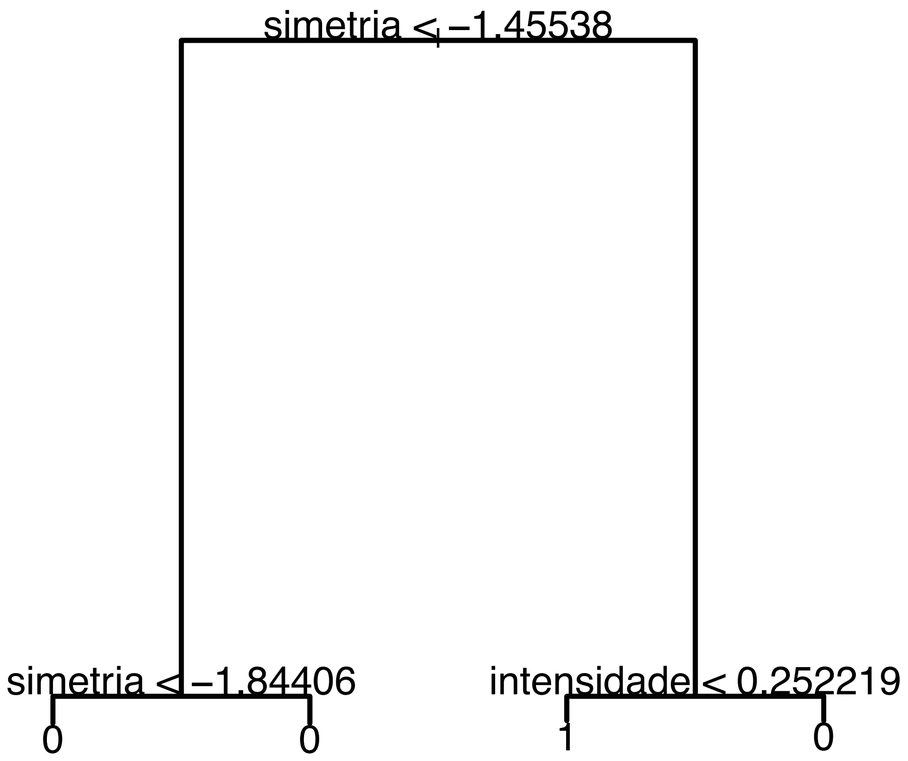
ou

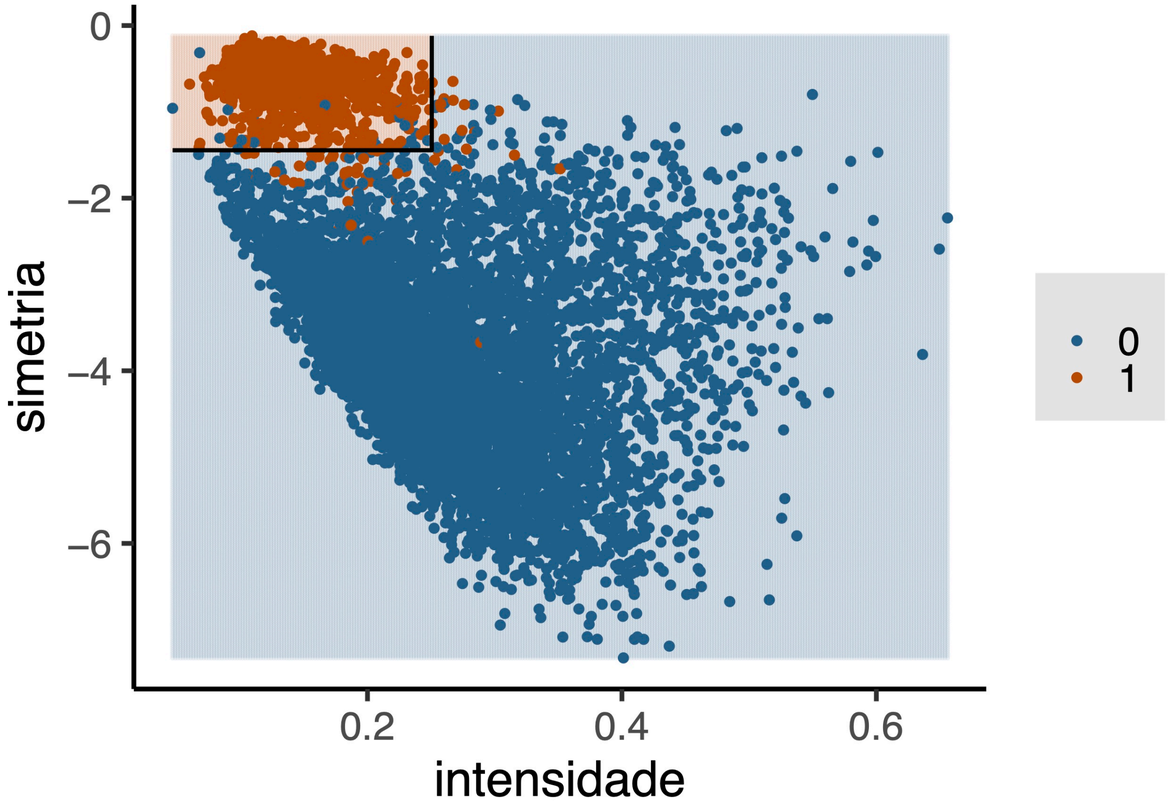
$$\widehat{E}_D(R_1, \ldots, R_J) = -\sum_{j=1}^{J} \sum_{k=1}^{K} \hat{p}_{jk} \log \hat{p}_{jk}$$

→ *entropia*

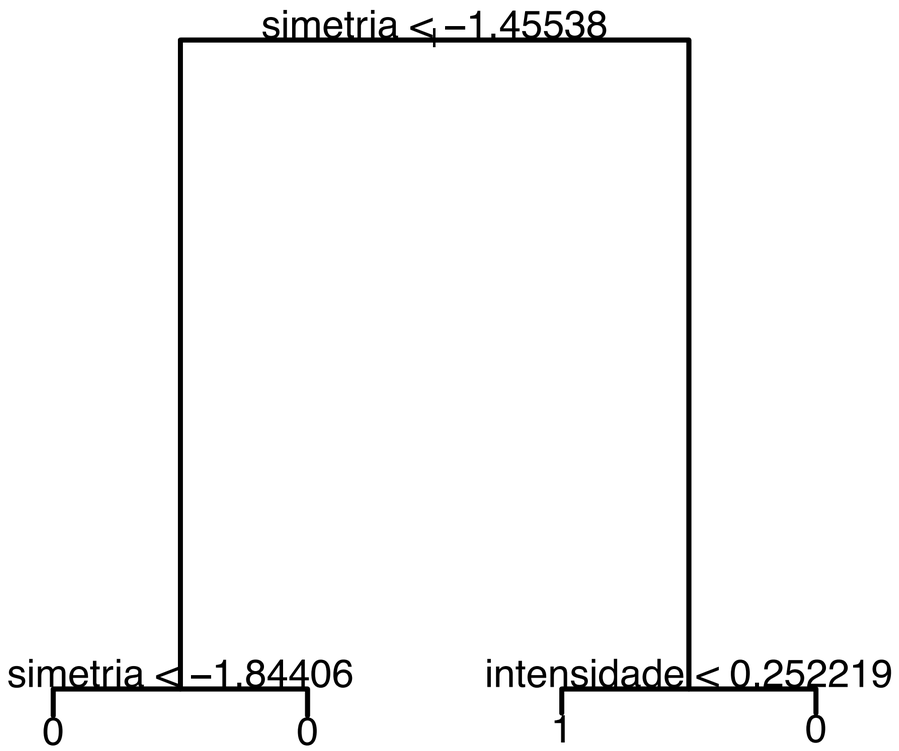


Árvores de classificação





Árvores de classificação



		Classe verdadeira	
		0	**1**
Classe predita	**0**	1736	34
	1	7	230

$$\text{Precisão} = \frac{1736 + 230}{2007} \times 100 = 98\,\%$$

Vantagens e desvantagens das árvores

- ➕ As árvores são muito fáceis de interpretar
- ➕ As árvores podem ser representadas graficamente
- ➕ As árvores podem utilizar misturas de variáveis quantitativas e qualitativas, sem necessidade de transformar as variáveis

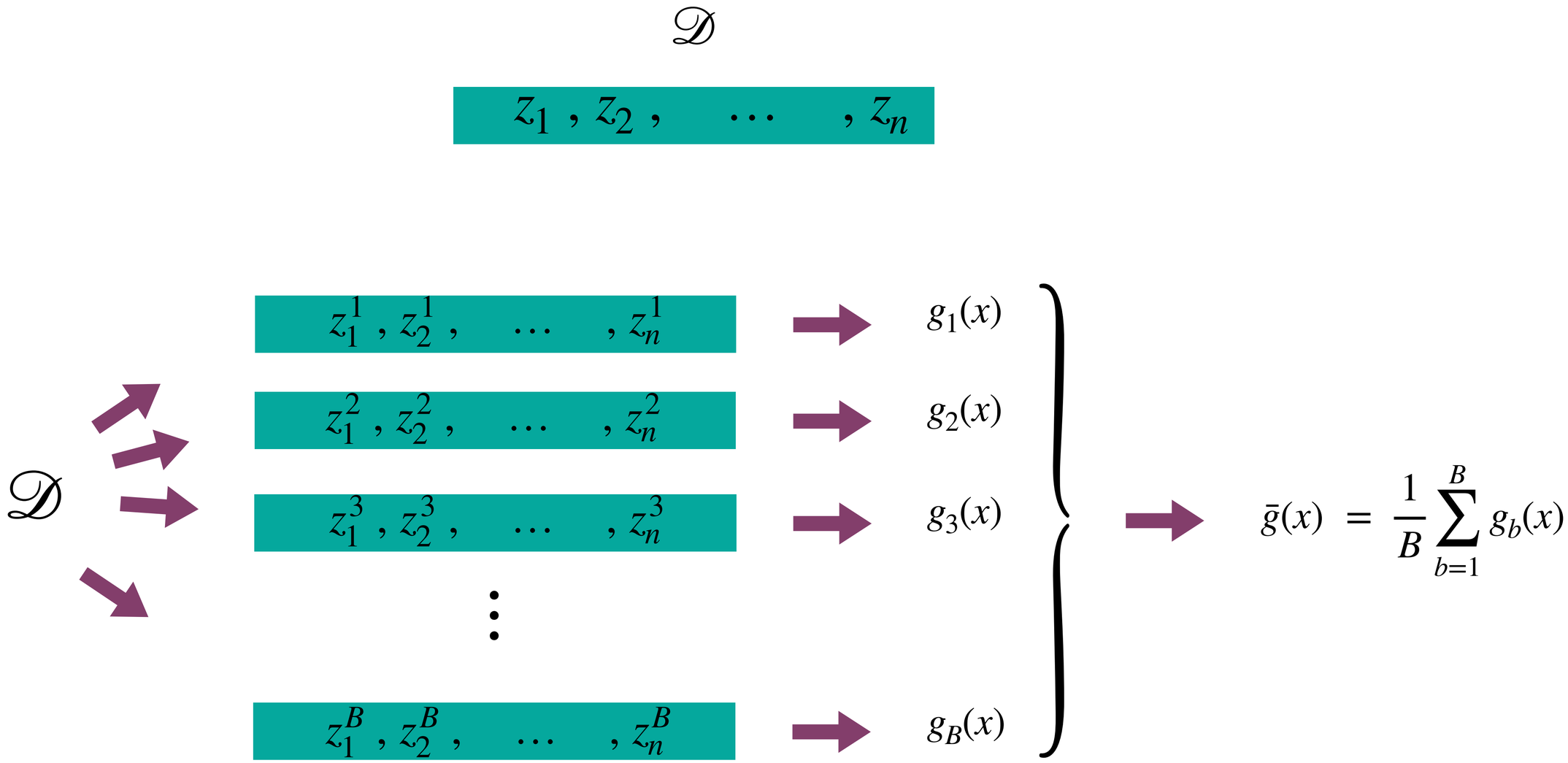
- ➖ Geralmente, as árvores não tem um alto grau de acurácia em comparação com outros métodos
- ➖ As árvores podem ser pouco robustas: pequenas perturbações nos dados podem causar grandes mudanças na árvore estimada

Bagging

- As árvores vistas até agora são modelos que em geral tem alta variância
- O método de *bootstrap aggregation*, ou *bagging,* é um procedimento geral para reduzir a variância de um método de aprendizagem estatística
- Lembremos que se temos n variáveis aleatórias independentes $Z_1, \ldots, Z_n$, cada uma com variância σ^2, a variância da média $\bar{Z}$ é σ^2/n. Então a média de um conjunto de observações tem o poder de *reduzir a variância*
- No caso de métodos preditivos, a ideia é construir B funções preditoras $g_1(x), \ldots, g_B(x)$ baseadas em amostras *bootstrap* extraídas do conjunto de dados $\mathscr{D}$ e construir um preditor dado pela média

$$\bar{g}(x) = \frac{1}{B}\sum_{b=1}^{B} g_b(x)$$

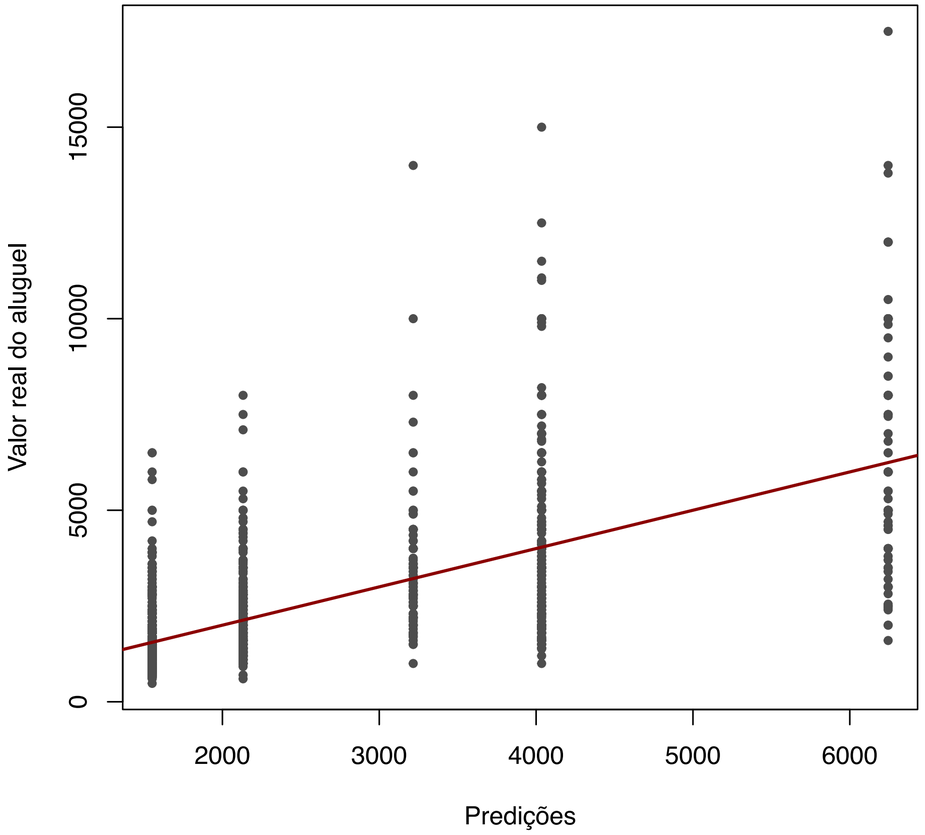
Bagging


$\mathscr{D}$
$z_1, z_2, \cdots, z_n$
$\mathscr{D}$
$z_1^1, z_2^1, \cdots, z_n^1$
$z_1^2, z_2^2, \cdots, z_n^2$
$z_1^3, z_2^3, \cdots, z_n^3$
$z_1^B, z_2^B, \cdots, z_n^B$
$g_1(x)$
$g_2(x)$
$g_3(x)$
$g_B(x)$
$\bar{g}(x) = \frac{1}{B}\sum_{b=1}^{B} g_b(x)$

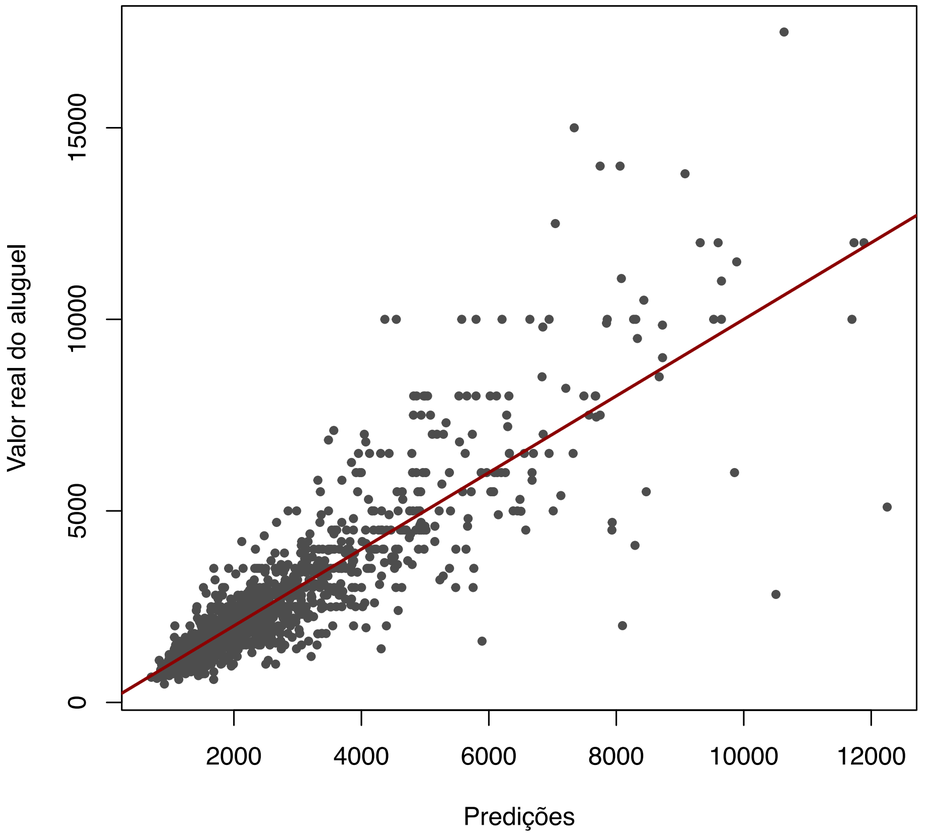
Bagging

- Para aplicar bagging para árvores de decisão, cada função $g_b(x)$ será a função de predição obtida por uma árvore de regressão ou classificação ajustada com os dados da amostra bootstrap $z_1^b, \ldots, z_n^b$
- Em geral, as árvores obtidas são deixadas com profundidade alta, sem ser podadas. Isto gera árvores com baixo viés e alta variância. A combinação destes preditores usando a média das predições reduz a variância do conjunto de cada preditor isolado.

Árvore de regressão



Bagging



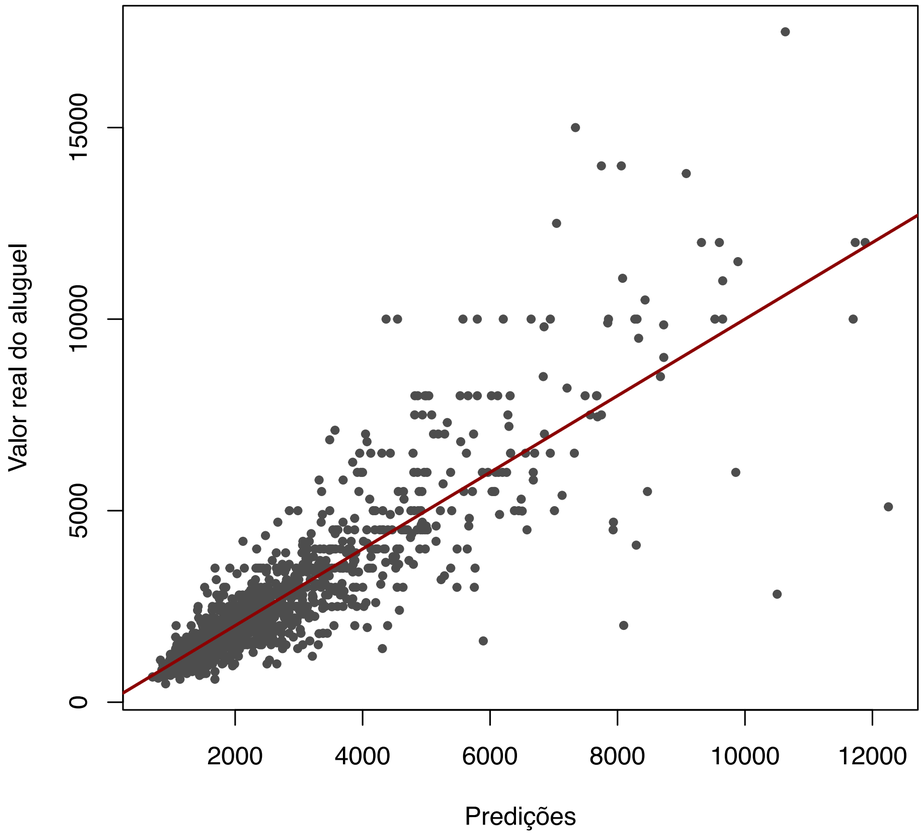
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 1543.43$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 976.64$$

Florestas aleatórias

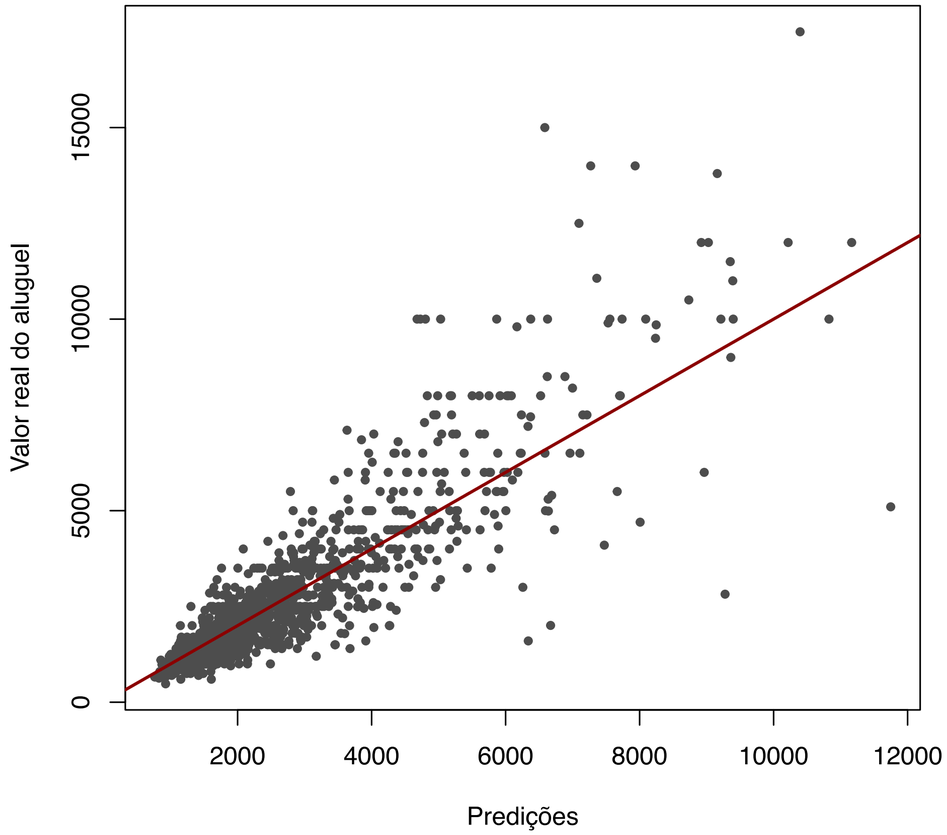
- Lembremos que a variância da média $\bar{Z}$ é σ^2/n somente no caso de variáveis $Z_1, \ldots, Z_n$ independentes! No caso de amostras bootstrap e suas funções derivadas, a hipótese de independência não é satisfeita e portanto a redução na variância pode ser consideravelmente menor.
- As florestas aleatórias tentam reduzir a dependência entre as funções $g_1(x), \ldots, g_B(x)$ utilizando subconjuntos de variáveis diferentes em cada divisão dos nós das árvores: cada vez que uma divisão vai ser feita, somente m variáveis preditoras são consideradas para definir a nova região (em vez das p variáveis disponíveis)
- Um novo subconjunto de preditoras é escolhido a cada nova divisão, e em geral é utilizado o valor $m \approx \sqrt{p}$

Bagging



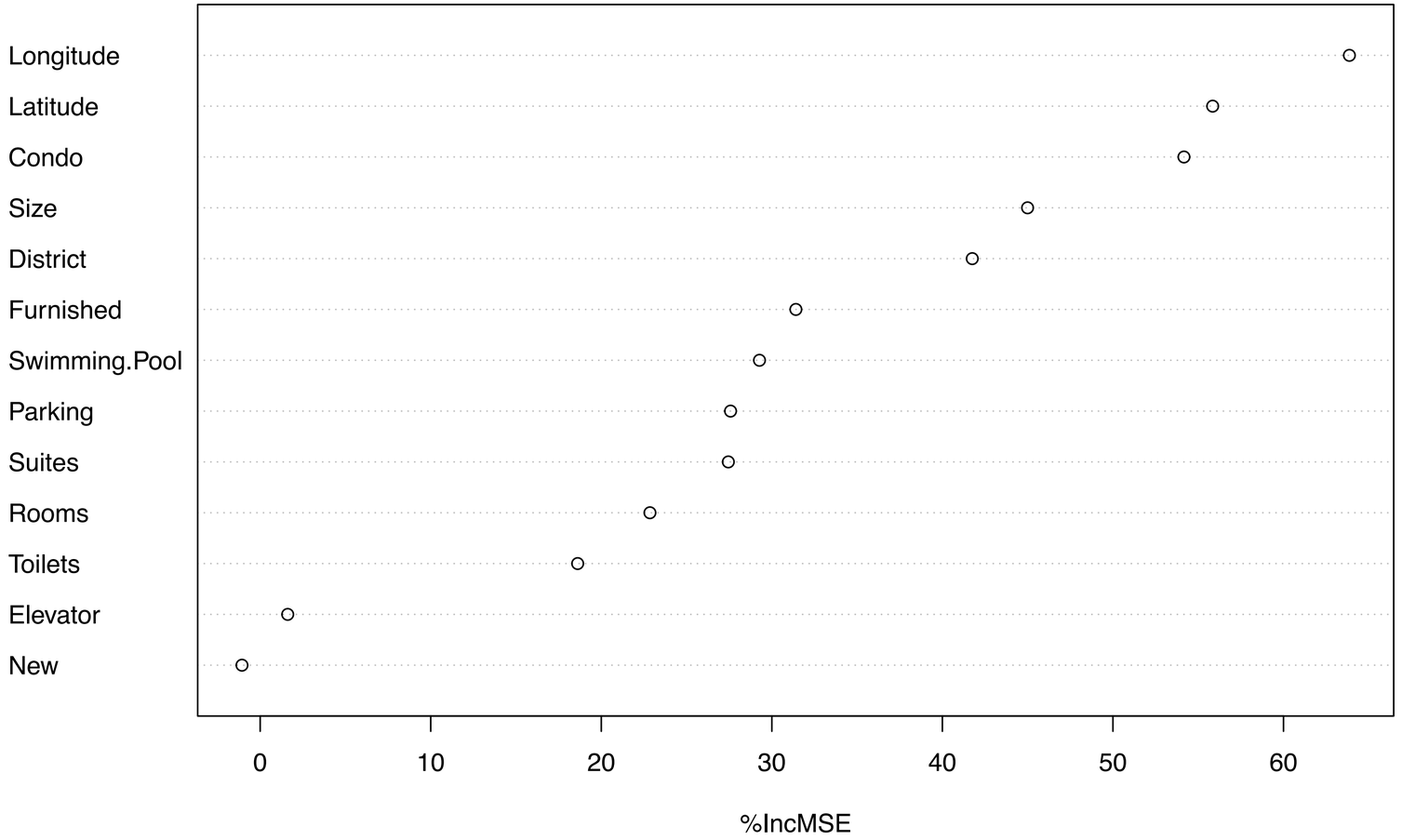
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 976.64$$

Florestas aleatórias



$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 985.19$$

Importância das variáveis



%IncMSE: para cada árvore, é calculado o erro de predição nos dados não incluídos na amostra bootstrap (out-of-bag). Logo, o mesmo é feito após permutar cada variável. %IncMSE é a média sobre todas as árvores da diferença entre os dois valores, para cada variável, normalizada pelo desvio padrão das diferenças.

Boosting para regressão

- Assim como o método de *bagging*, o método *boosting* é um método geral que pode ser aplicado a diferentes modelos e técnicas de aprendizagem estatística para regressão ou classificação
- O método boosting é similar a *bagging* no sentido que a função preditora final é obtida a partir da soma de preditoras $g_1(x), \ldots, g_B(x)$. Mas em vez de obter $g_i(x)$ em função de uma amostra bootstrap, ela é obtida num conjunto de dados tendo como variáveis preditoras as variáveis preditoras originais de $\mathscr{D}$ e como variáveis resposta, os resíduos $y_i - g_{i-1}(x)$

Boosting para regressão

O resultado é uma função $g(x) = \lambda \sum_{b=1}^{B} g_b(x)$ obtida como uma soma das funções preditoras

$g_1(x), \ldots, g_B(x).$

A taxa λ é chamada de *taxa de aprendizagem,* e controla o peso relativo de cada modelo individual

$g_i(x), i = 1,\ldots B$

Quanto menor o valor de λ, maior deve ser o valor de B para obtermos uma mesma acurácia, e vice-versa.

Algoritmo: *boosting* para árvores de regressão

1. Inicialize $\bar{g}(x) = 0$ e $r_i^1 = y_i$ para todo i no conjunto de treinamento
2. Para $b = 1, \ldots, B$, repita:
   a. Ajuste uma árvore $g_b(x)$ com $d + 1$ folhas para os dados de treinamento $\mathscr{D}_b = \{(x_1, r_1^b), \ldots, (x_n, r_n^b)\}$

   b. Atualize $\bar{g}(x)$ adicionando a função escalada $g_b(x)$, isto é faça

$$\bar{g}(x) \leftarrow \bar{g}(x) + \lambda g_b(x)$$

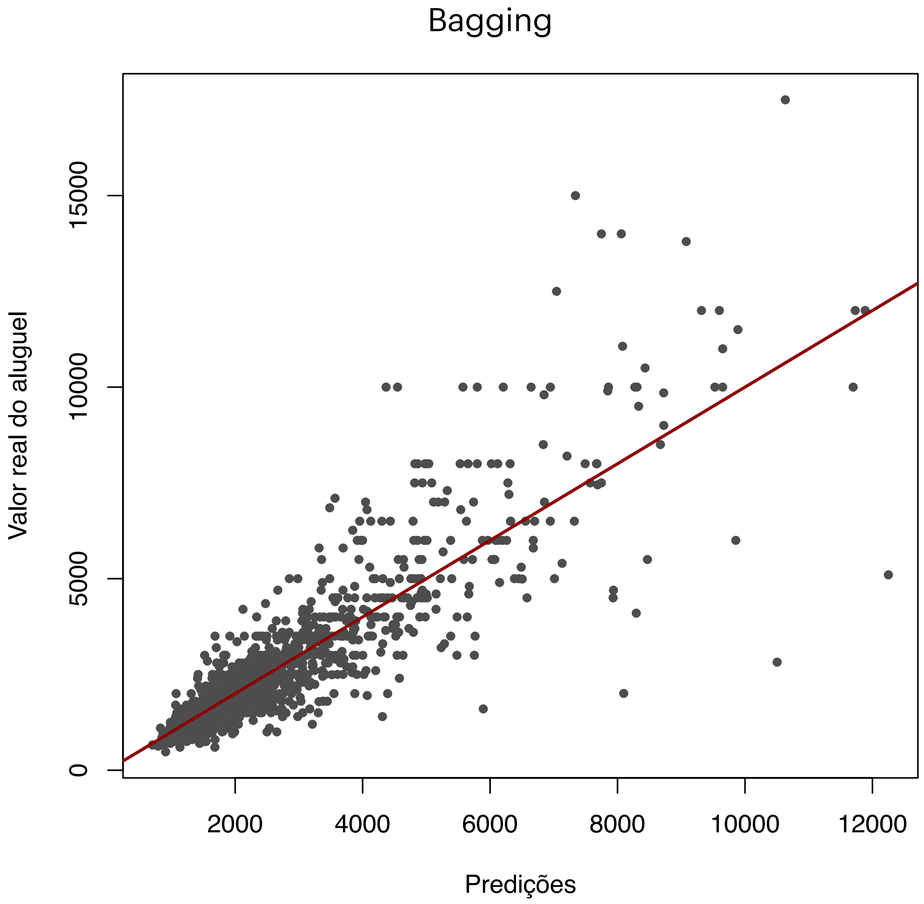
   c. Atualize os resíduos $r_i^{b+1} = r_i^b - \lambda g_b(x_i)$

3. Devolva o modelo final $\bar{g}(x) = \lambda \sum_{b=1}^{B} g_b(x)$

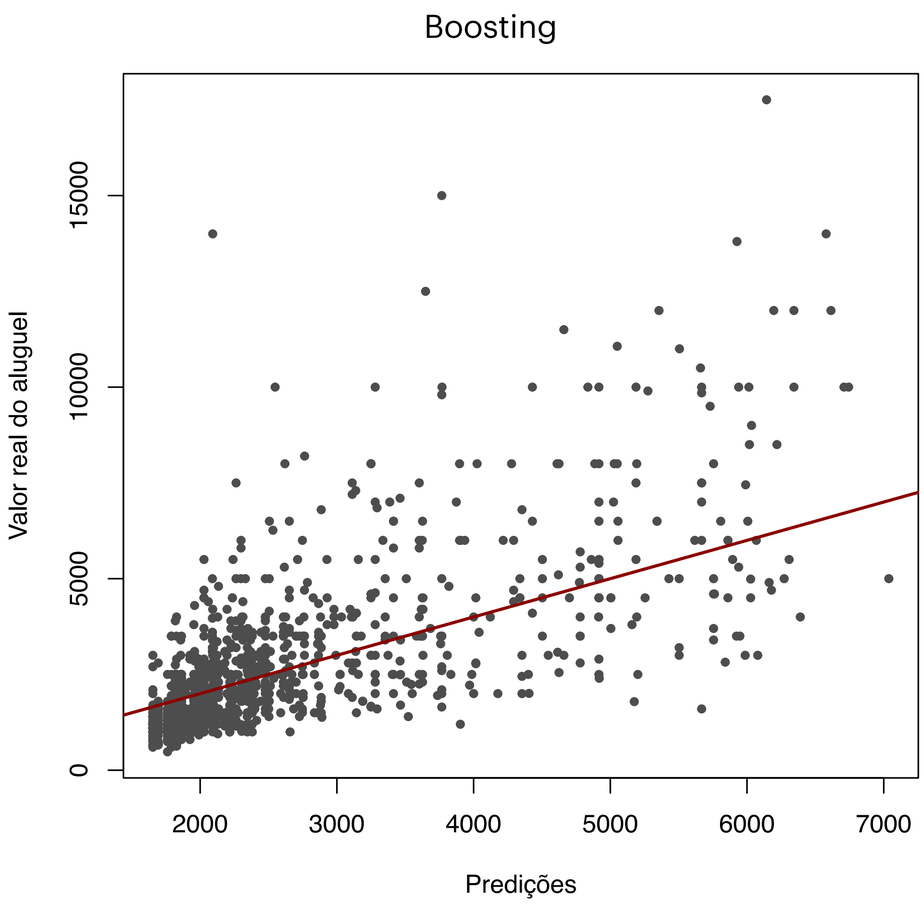
Ajuste dos parâmetros de suavizado

Boosting tem três parâmetros que devem ser definidos:

- O número de árvores B: dependendo do valor, o modelo pode ter maior ou menor viés e variância. Este valor pode ser escolhido por validação cruzada.
- A parâmetro de escala λ: controla o impacto de cada árvore no modelo final, valores típicos são pequenos como 0.01 ou 0.001. Está relacionado com o valor de B e também pode ser escolhido por validação cruzada.
- O número de divisões permitidas em cada árvore d: determina a interação entre as variáveis no modelo. Se $d = 1$ não há interação e o modelo é um modelo aditivo onde cada termo depende de uma única variável.

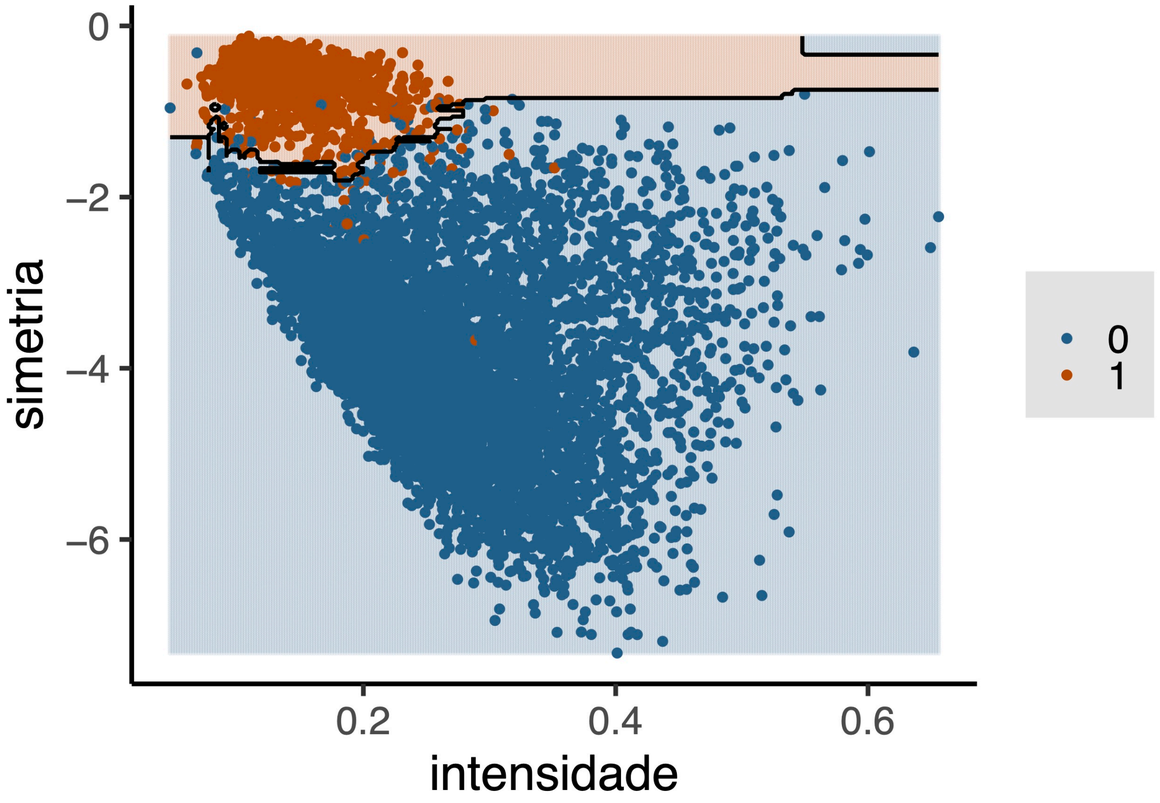


$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 976.64$$



$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 1466.76$$

Boosting para classificação



		Classe verdadeira	
		0	**1**
Classe predita	**0**	1734	31
	1	9	233

$$\text{Precisão} = \frac{1734 + 233}{2007} \times 100 = 98\,\%$$