Interpretando Modelos de Regressão Simbólica

Guilherme Seidyo Imai Aldeia
Orientador: Prof. Dr. Fabrício Olivetti de França

Universidade Federal do ABC
Programa de Pós Graduação em Ciência da Computação
Heuristics, Analysis and Learning Laboratory (HAL)

Santo André/SP
31 de julho de 2020

Índice

Modelos interpretáveis e não interpretáveis

- Modelos de regressão tradicionais: aprendizagem estatística, modelagem matemática
  - Interpretação do efeito das variáveis pode ser obtida
  - Plausibilidade pode ser verificada
  - Explicabilidade pode ser analisada
- Modelos complexos: baseados em hiperplanos, instância, bioinspirados
  - Como interpretar?
  - Como ver se o modelo é plausível?
  - Ele explica algum modelo teórico?

Modelos interpretáveis e não interpretáveis

Modelo caixa preta

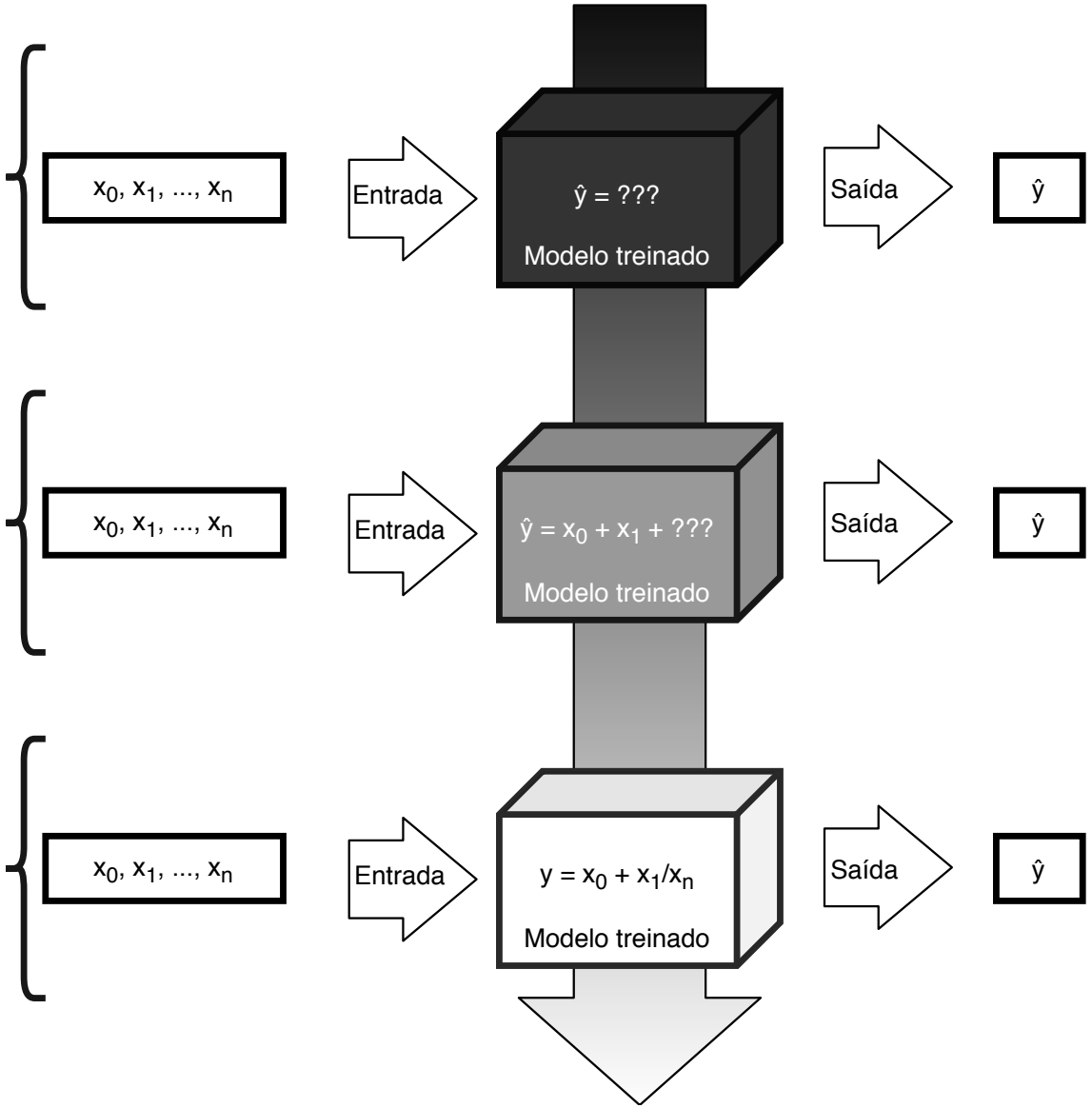
Não é possível inspecionar o modelo para obter qualquer insight sobre seu funcionamento. Sabemos a entrada e a saída, mas não os processos internos

Modelo caixa cinza

O modelo é parcialmente interpretável, onde parte dos símbolos tem significado explícito. Pode ser resultante da combinação de um modelo caixa branca e outro caixa preta

Modelo caixa branca

Podemos entender o funcionamento do modelo. Lembrando que existe subjetividade no entendimento do resultado de um modelo



Model Agnostics

- Ferramentas para interpretar modelos caixa preta
- LIME, ELI5, SHAP...
- SHAP:
  - Baseado em teoria dos jogos
  - Busca aproximar a importância de uma variável comparando o desempenho do regressor com e sem ela
  - Explicação é boa?

Regressão simbólica

- Regressão simbólica - alternativa caixa cinza
  - Implementada por meios de Computação Evolutiva
  - Busca por funções matemáticas
- Representação das soluções
  - Tradicionalmente feita por árvores
  - Representação IT: mais restritiva

Índice

Características

- Forma de representar soluções para a regressão simbólica, que:
  - Restringe o espaço de busca
  - Prioriza funções simples
- Constrói funções à partir de Termos IT:
  - Termos que descrevem a aplicação de uma função de transformação sobre interações das variáveis originais

Termo IT

Seja a interação das n variáveis do problema dada por:

$$i(\mathbf{x}) = \prod_{i=1}^{n} x_i^{k_i},$$

um **Termo IT** é uma tupla:

$$(t, \mathbf{k}),$$

com $t : \mathbb{R} \to \mathbb{R}$, $\mathbf{k} \in \mathbb{R}^n$. Dessa forma, a tupla constrói a composição:

$$it(\mathbf{x}) = (t \circ i)(\mathbf{x}),$$

utilizando o vetor de expoentes $\mathbf{k}$ para calcular a interação.

Expressão IT

Uma **Expressão IT** é a combinação linear de m Termos IT:

$$ITExpr(\mathbf{x}) = \sum_{j=1}^{m} w_j \cdot it_j(\mathbf{x}) + b = \sum_{j=1}^{m} w_j \cdot (t_j \circ i)(\mathbf{x}) + b$$

Exemplo

Sejam os termos t_1 e t_2 para um problema de duas variáveis dado por:

$$\begin{cases} t_1 = (id, [1, 0]) \\ t_2 = (cos, [0, -1]) \end{cases}$$

a expressão equivalente é:

$$ITExpr(x_1, x_2) = w_1 \cdot id(x_1^1 \cdot x_2^0) + w_2 \cdot cos(x_1^0 \cdot x_2^{-1}) + b,$$

$$\Rightarrow ITExpr(x_1, x_2) = w_1 \cdot x_1^1 + w_2 \cdot cos(\frac{1}{x_2}) + b,$$

onde w_1, w_2, b são coeficientes ajustados por meio de modelos de regressão lineares.

Exemplos representáveis e não representáveis

Expressões representáveis	Expressões não representáveis
$x_1 + \pi * cos(x_2)$	$\sqrt{x_1 * x_1} + tanh(tanh(tanh(x_2)))$
$x_1 + x_2 + x_1 * x_2$	$\frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}$
$10 * log(\frac{I}{I_0}) + 5.0$	$\frac{sen(x_1 * x_2)}{x_3}$
$a^2 + 2 * a * b + b^2$	$\sqrt{a^2 + b^2}$

Índice

Descrição

- Estratégia evolutiva baseada em mutação que utiliza a representação IT para implementar a regressão simbólica
- Implementação em *Python*
- *Fitness* calculado pelo RMSE

Mutações

- **Add:** Adiciona um Termo IT
- **Drop:** Remove um Termo IT
- **Term:** Modifica os expoentes de um termo
- **Func:** Modifica a função de transformação de um termo
- **Positive interaction:** Adiciona um Termo IT resultante da interação positiva entre 2 termos já existentes
- **Negative interaction:** Adiciona um Termo IT resultante da interação negativa entre 2 termos já existentes

Pseudocódigo

Algoritmo 1: *Interaction-Transformation Evolutionary Algorithm* (ITEA)

Result: Função simbólica $\widehat{f}$

pop $\leftarrow$ [n expressões aleatórias];

for *g gerações* **do**

  pop $\leftarrow$ pop $+$ mutação(pop);

  pop $\leftarrow$ torneio(pop) ;

end

Índice

Intuição

- Medir o efeito médio que uma mudança em uma variável causa no resultado da função
  - Calculado utilizando a derivada da expressão - interpretação geométrica é a taxa de variação
  - Modelos complexos: derivada difícil de ser calculada, obtida apenas de forma numérica, ou não existente.

Average Marginal Effects (AME)

- AME da variável *i*: derivar parcialmente em termos de i, avaliar para todos os dados, calcular a média.
  - Obter o valor médio da contribuição de cada variável.
  - Absorve a variância entre as observações
  - Valor único representando a variável

Marginal Effects at the Means (MEM)

- MEM: derivar parcialmente em termos de i, fixar as covariáveis no valor médio, obter o efeito marginal variando i em um intervalo
  - Analisa contribuição de cada variável ao longo de um intervalo, enquanto as outras variáveis são fixadas na média
  - Útil quando há relações complexas entre as variáveis Visualização gráfica

Expressão IT - derivadas

Regressão ITEA: expressões encontradas podem ser derivadas facilmente por um procedimento algorítmico, conhecendo a derivada de t:

$$\frac{\partial ITExpr(\mathbf{x})}{\partial x_i} = w_1 \cdot g_1'(\mathbf{x}) + \ldots + w_m \cdot g_m'(\mathbf{x}),$$

onde

$$g'(\mathbf{x}) = (t_i' \circ i)(\mathbf{x}) \cdot i'(\mathbf{x}),$$

$$i'(\mathbf{x}) = k_i \frac{i(\mathbf{x})}{x_i}.$$

Índice

Design dos experimentos

Teste para base sintética e base de dados do mundo real de baixa dimensionalidade

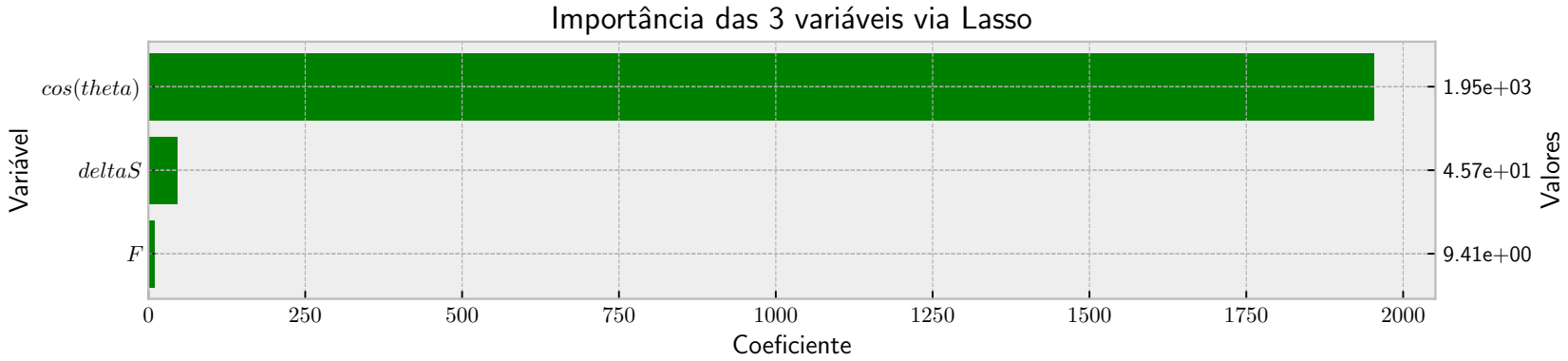
- Comparar o efeito marginal e as explicações pelo SHAP para uma solução do ITEA
- Comparar explicações nativas do XGB para o mesmo problema, e também analisá-lo com o SHAP
- Comparar os resultados do ITEA e XGB em termos da interpretação e explicação obtidas

Base sintética: *"Sanity Check"*

- Testar o método para uma expressão conhecida e representável
- Objetivo: assegurar de que não há bugs ou erros de implementação
- Equação sem ruído e fácil interpretação - Trabalho:

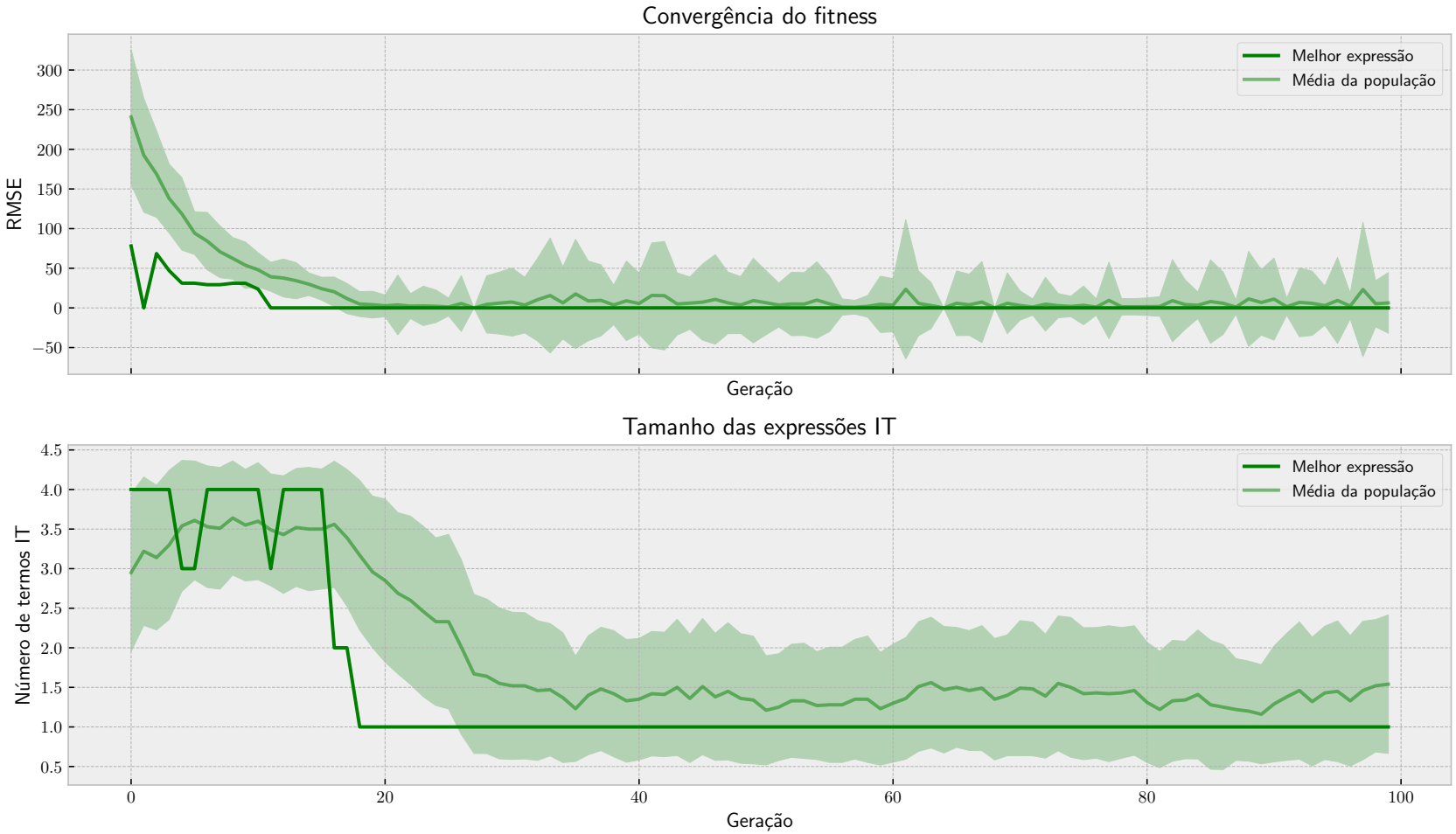
$$W = F \cdot \Delta s \cdot cos(\theta)$$

"Sanity Check" - inspeção do *dataset*



	F	deltaS	cos(theta)	W
count	100.000000	100.000000	100.000000	100.000000
mean	101.346925	19.539549	0.481168	937.500255
std	19.034768	5.207653	0.153864	397.216711
min	35.536656	6.987805	0.127898	197.881509
25%	85.948749	15.781728	0.353948	650.736746
50%	102.717544	20.191799	0.490962	901.375186
75%	113.835432	22.951400	0.576644	1200.111696
max	144.515978	31.845599	0.792238	1996.435894

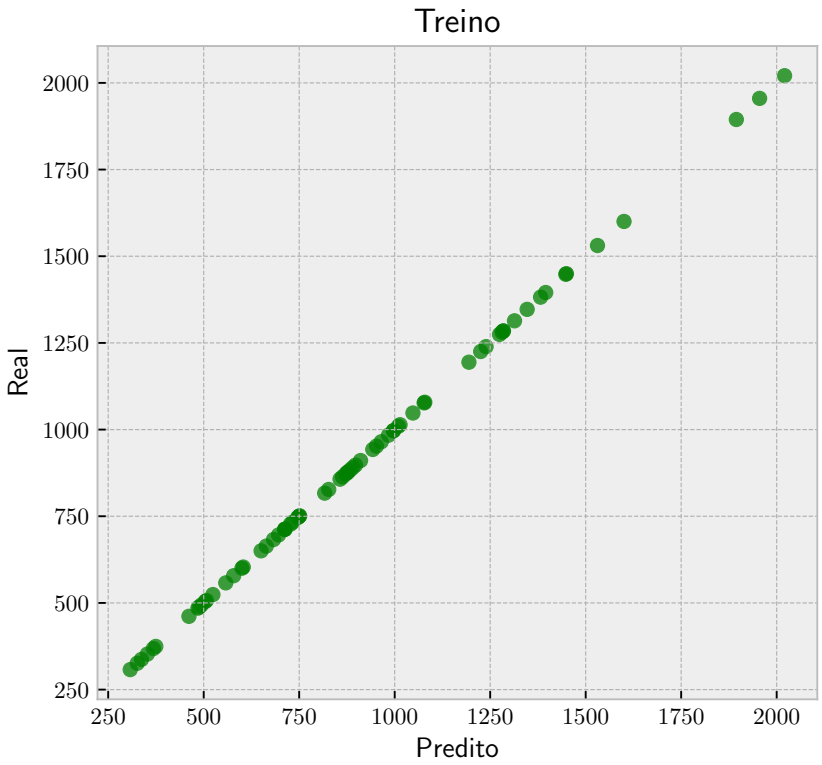
"Sanity Check"- Execução ITEA



"Sanity Check" - Execução ITEA

	Valor
Tempo (ms)	5.564491
Score final (RMSE)	0.000000
Tamanho final (termos IT)	1

	Treino	**Teste**
R2	1.000000	1.000000
R2 ajustado	1.000000	1.000000
MSE	0.000000	0.000000
n observações	70	30
Pearson rho	1.000000	1.000000
Pearson 2-tail p	0.000000	0.000000



"Sanity Check" - Execução ITEA

Coeff	Coeff SE	Coeff 95% IC	Func	Strengths	Term
1.0	0.0	0.0	id	[1, 1, 1]	$1.0 \cdot F \cdot deltaS \cdot cos(theta)$
0.0	0.0	0.0	(intercept)	—	0.0

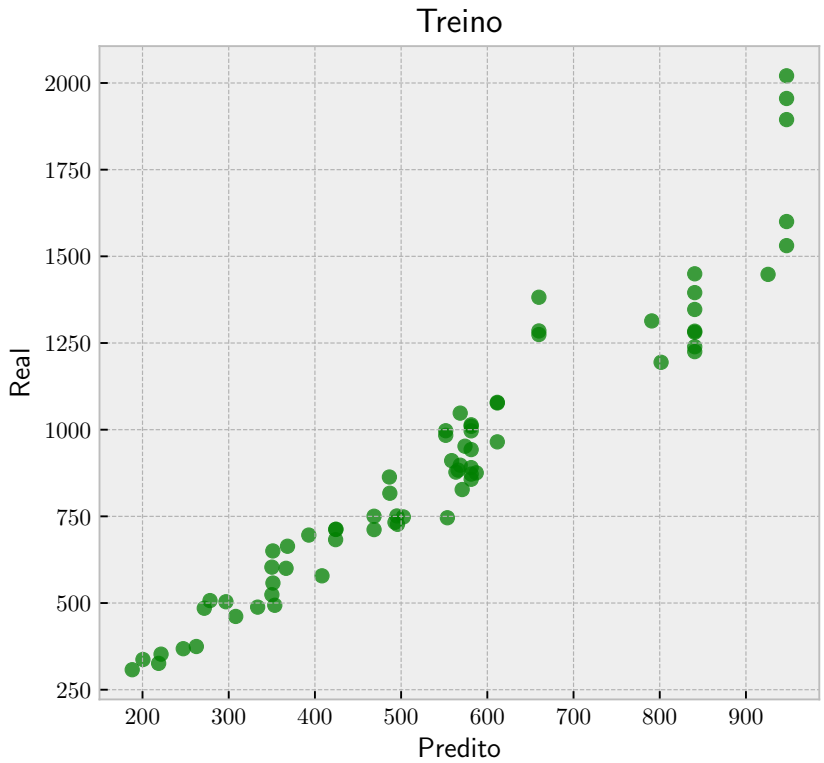
Expressão encontrada e suas derivadas:

$$\text{ITExpr} = \underbrace{F \cdot deltaS \cdot cos(theta)}_{\text{termo 0}} + 0.0$$

$$\begin{cases} \frac{\partial ITExpr}{\partial F} = \underbrace{deltaS \cdot cos(theta)}_{\text{termo 0}} \\ \frac{\partial ITExpr}{\partial deltaS} = \underbrace{F \cdot cos(theta)}_{\text{termo 0}} \\ \frac{\partial ITExpr}{\partial cos(theta)} = \underbrace{F \cdot deltaS}_{\text{termo 0}} \end{cases}$$

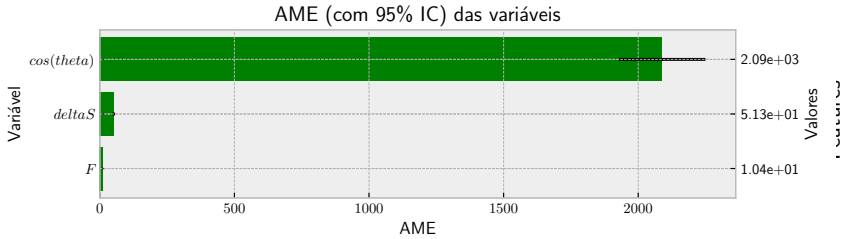
"Sanity Check" - Execução XGB

	Treino	**Teste**
R2	-0.128493	-0.510974
MSE	170963.796875	3935959.125000
n observações	70	30
Pearson rho	0.969387	0.852146
Pearson 2-tail p	0.000000	0.000000

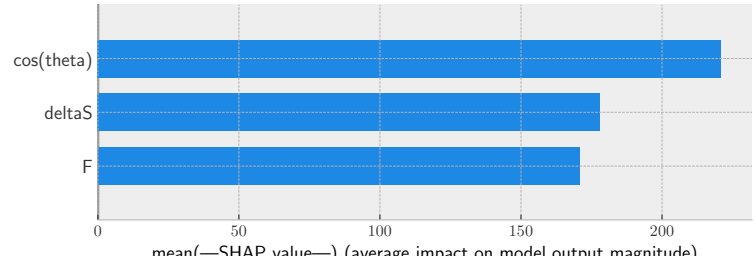


"Sanity Check" - Importâncias

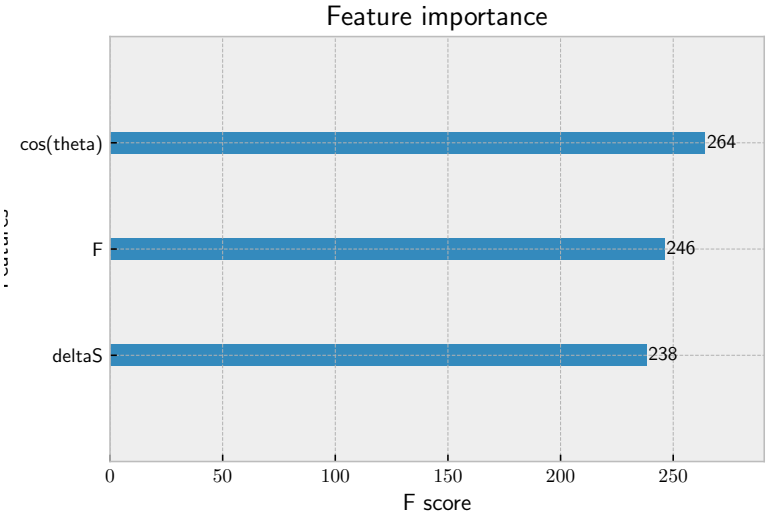
AME-ITEA



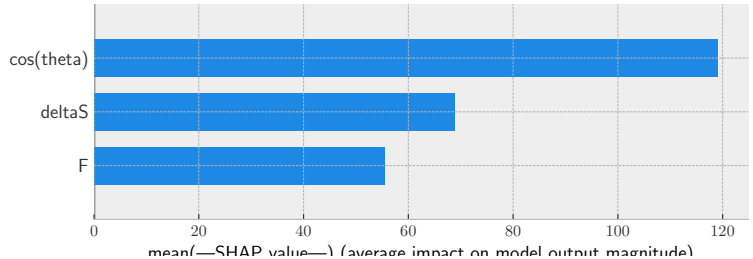
SHAP-ITEA



FScore-XGB

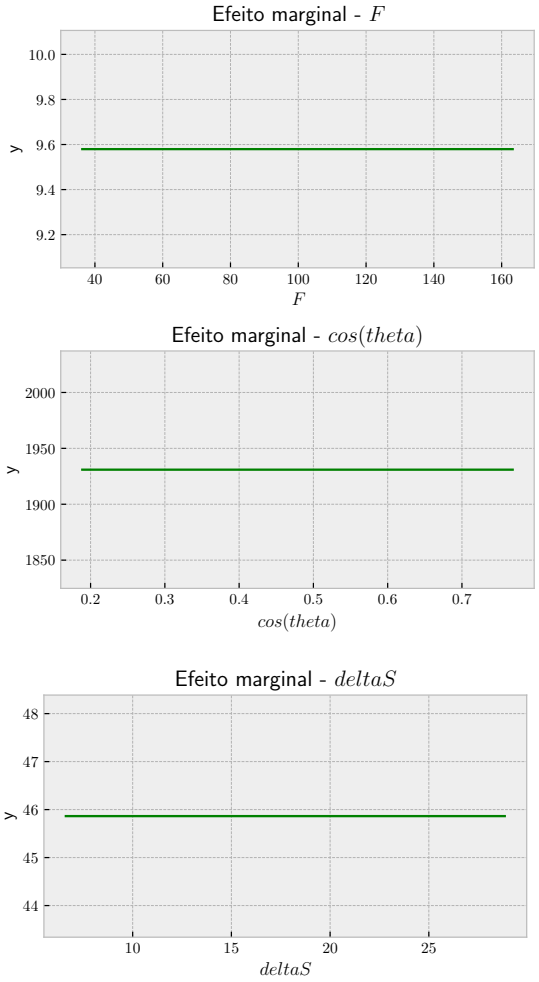


SHAP-XGB

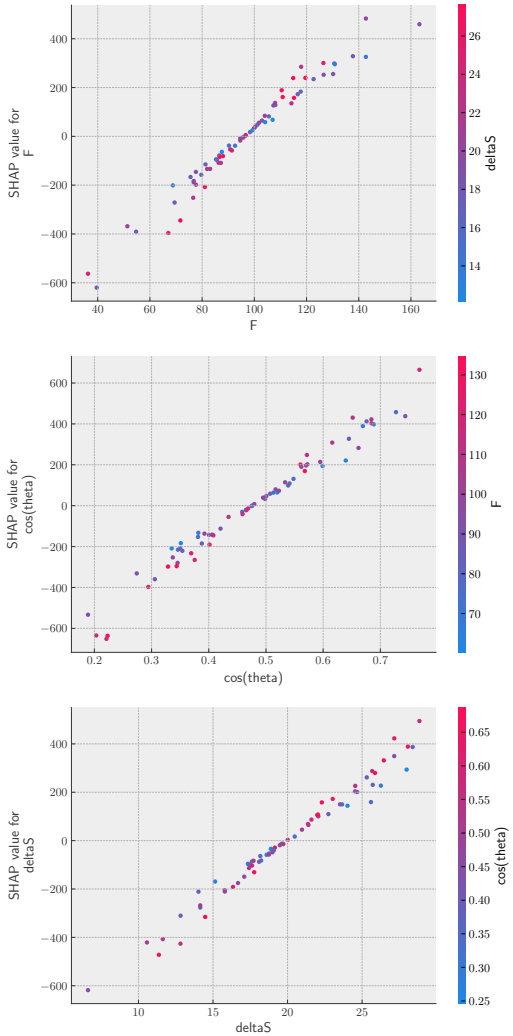


"Sanity Check" - Contribuição das variáveis

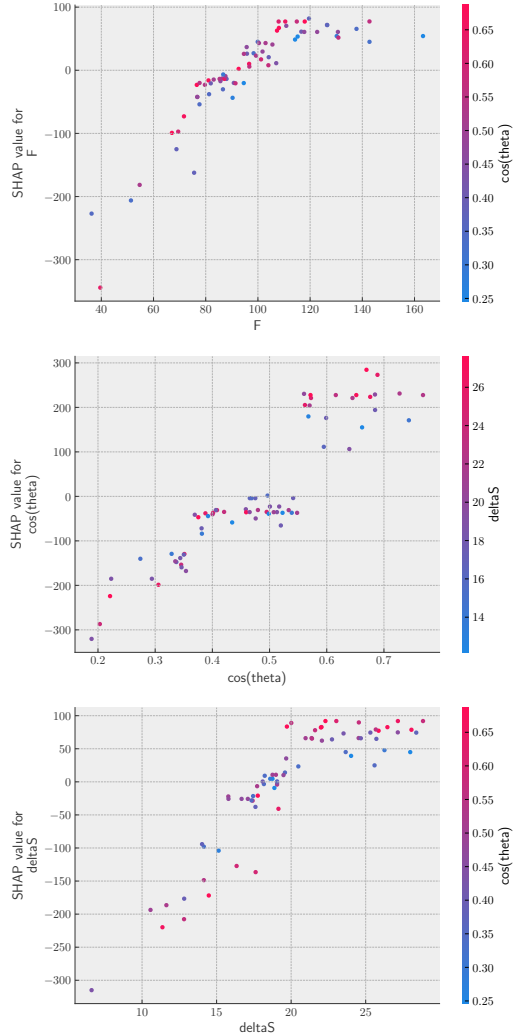
MEM-ITEA



SHAP-ITEA

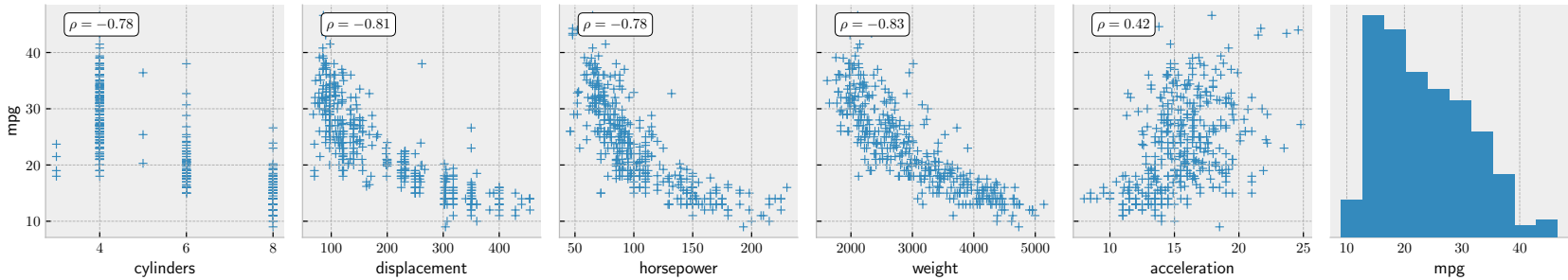


SHAP-XGB

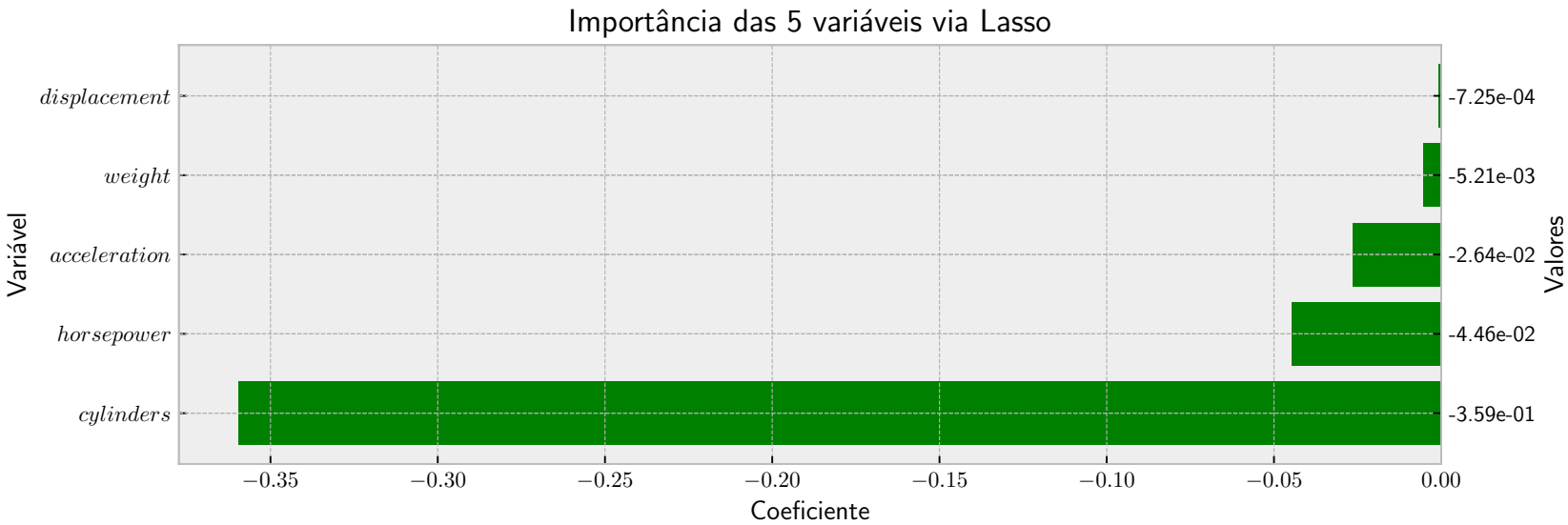


Base real: *Cars Dataset*

- Tentar predizer o consumo (miles per gallon, mpg) à partir de informações do motor:
  - n. de cilindros, cilindrada, cavalos de potência, peso, aceleração

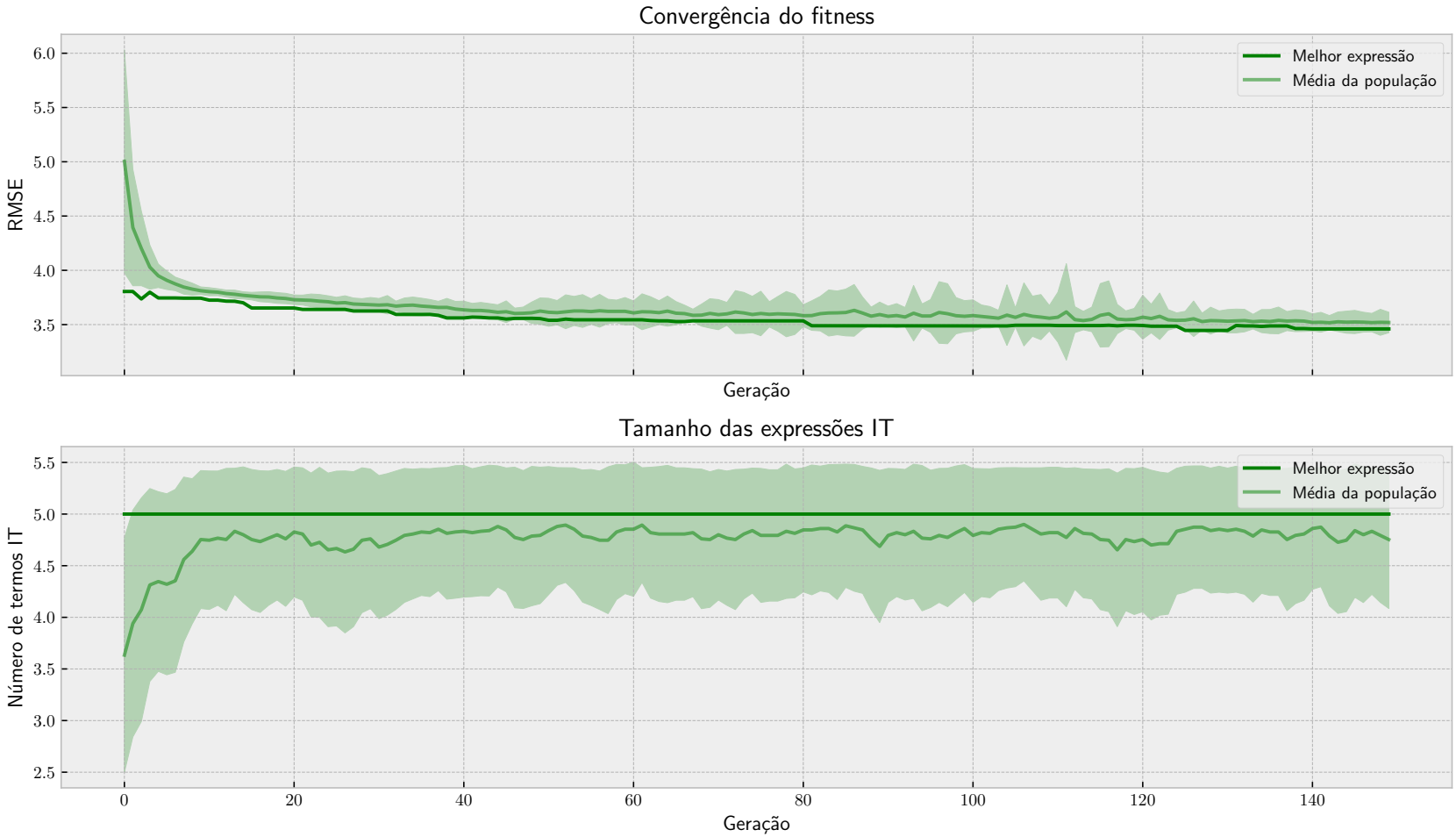


Cars Dataset - inspeção do *dataset*



	cylinders	displacement	horsepower	weight	acceleration	mpg
count	391.000000	391.000000	391.000000	391.000000	391.000000	391.000000
mean	5.475703	194.661125	104.511509	2979.751918	15.543990	23.436829
std	1.706337	104.661608	38.531429	849.404444	2.761894	7.812930
min	3.000000	68.000000	46.000000	1613.000000	8.000000	9.000000
25%	4.000000	105.000000	75.000000	2227.000000	13.750000	17.000000
50%	4.000000	151.000000	94.000000	2807.000000	15.500000	22.500000
75%	8.000000	284.500000	127.000000	3616.500000	17.050000	29.000000
max	8.000000	455.000000	230.000000	5140.000000	24.800000	46.600000

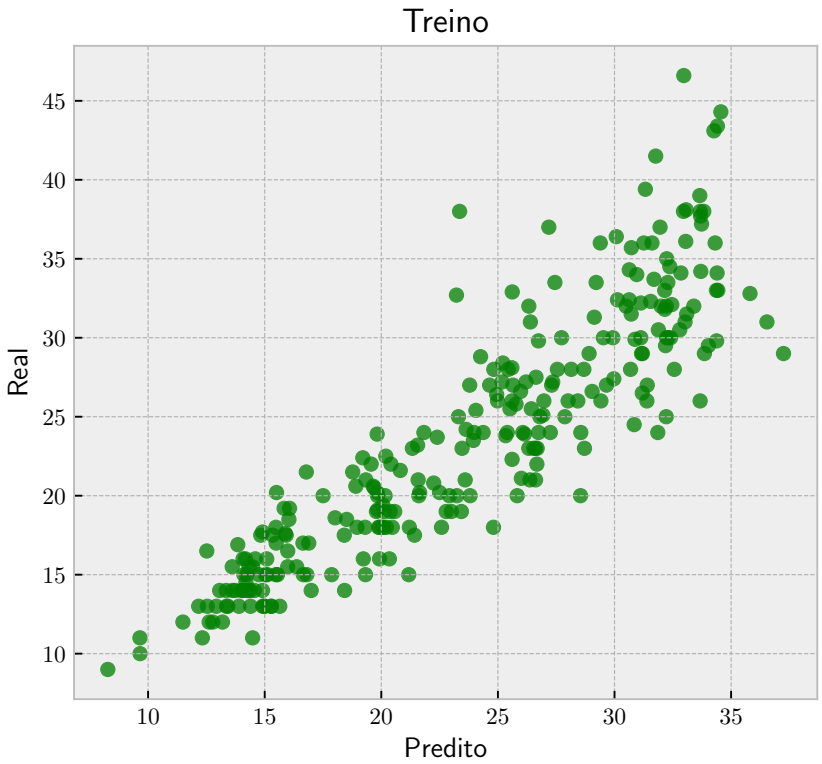
Cars Dataset - Execução ITEA



Cars Dataset - Execução ITEA

	Valor
Tempo (ms)	209.219749
Score final (RMSE)	3.460106
Tamanho final (termos IT)	5.000000

	Treino	**Teste**
R2	0.804073	0.750637
R2 ajustado	0.800403	0.739504
MSE	11.972335	14.952710
n observações	273	118
Pearson rho	0.915934	0.903541
Pearson 2-tail p	0.000000	0.000000



Cars Dataset - Execução ITEA

Coeff	**Coeff SE**	**Coeff 95% CI**	**Func**	**Strengths**	IT term
148.690653	1.060286	0.126568	cos	[3 1 -3 0 0]	$148.69 \cdot cos(\frac{cylinders^3 \cdot displacement}{horsepower^3})$
-0.000614	0.083226	0.009935	sqrt	[-2 -2 3 2 1]	$-0.0 \cdot \sqrt{\frac{horsepower^3 \cdot weight^2 \cdot acceleration}{cylinders^2 \cdot displacement^2}}$
-2.735990	0.072285	0.008629	log	[3 3 1 -2 1]	$-2.74 \cdot log(\frac{cylinders^3 \cdot displacement^3 \cdot horsepower}{acceleration \cdot weight^2})$
-3.789147	0.375106	0.044777	sin	[1 3 -1 -2 1]	$-3.79 \cdot sin(\frac{cylinders \cdot displacement^3 \cdot acceleration}{horsepower \cdot weight^2})$
-4.356428	0.299619	0.035766	sin	[-1 3 1 -3 2]	$-4.36 \cdot sin(\frac{displacement^3 \cdot horsepower}{acceleration^2 \cdot cylinders \cdot weight^2})$
-79.869684	1.978291	0.236151	(intercept)	-	-79.87

Expressão encontrada:

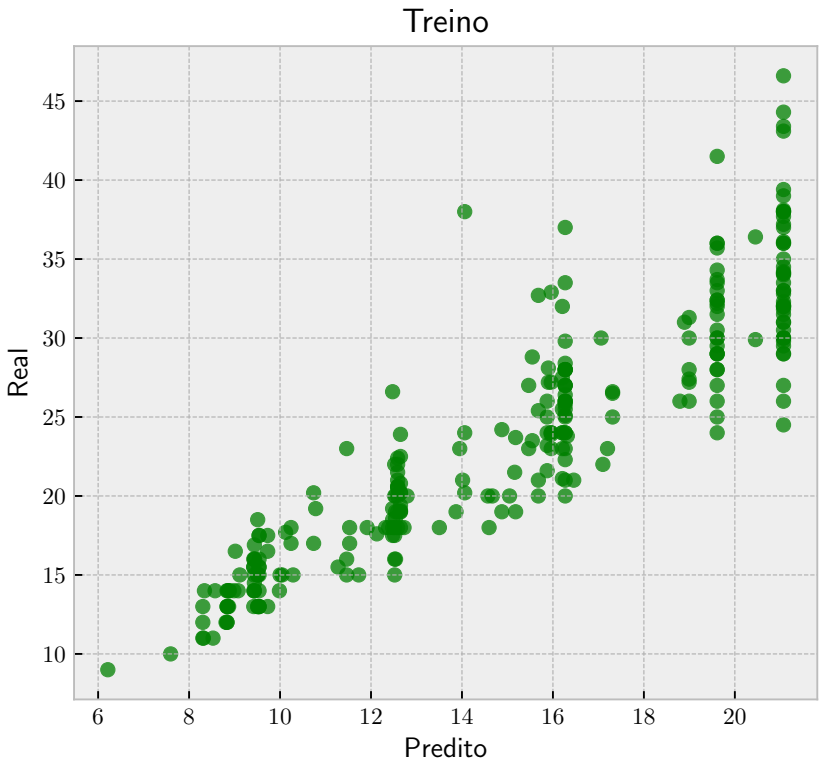
$$ITExpr = \underbrace{148.69 \cdot cos(\frac{cylinders^3 \cdot displacement}{horsepower^3})}_{\text{termo 0}} + \underbrace{-0.0 \cdot \sqrt{\frac{horsepower^3 \cdot weight^2 \cdot acceleration}{cylinders^2 \cdot displacement^2}}}_{\text{termo 1}} +$$

$$\underbrace{-2.74 \cdot log(\frac{cylinders^3 \cdot displacement^3 \cdot horsepower}{acceleration \cdot weight^2})}_{\text{termo 2}} + \underbrace{-3.79 \cdot sin(\frac{cylinders \cdot displacement^3 \cdot acceleration}{horsepower \cdot weight^2})}_{\text{termo 3}} +$$

$$\underbrace{-4.36 \cdot sin(\frac{displacement^3 \cdot horsepower}{acceleration^2 \cdot cylinders \cdot weight^2})}_{\text{termo 4}} + -79.87$$

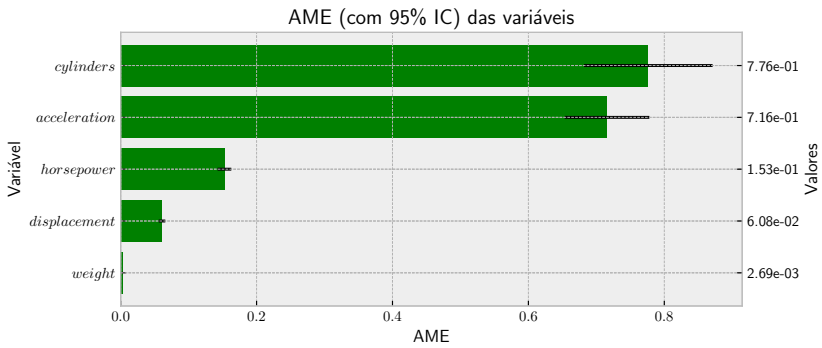
Cars Dataset - Execução XGB

	Treino	**Teste**
R2	-0.529839	-0.421525
MSE	93.482285	85.239670
n observações	273.000000	118.000000
Pearson rho	0.928236	0.903623
Pearson 2-tail p	0.000000	0.000000

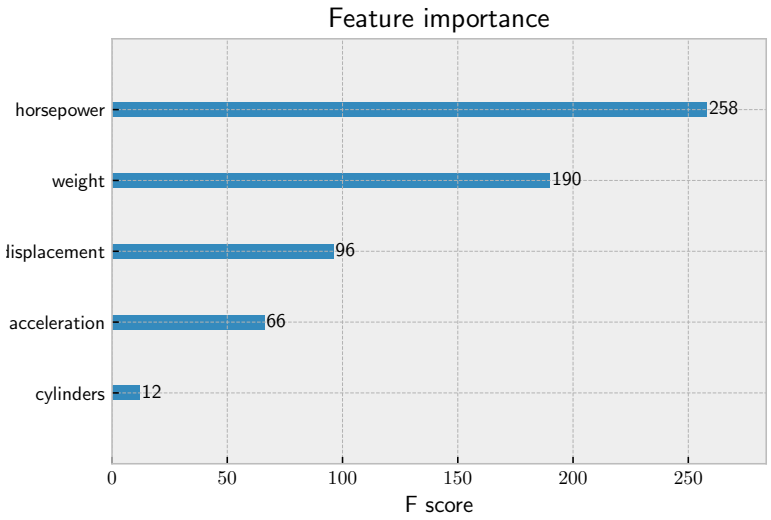


Cars Dataset - Importâncias

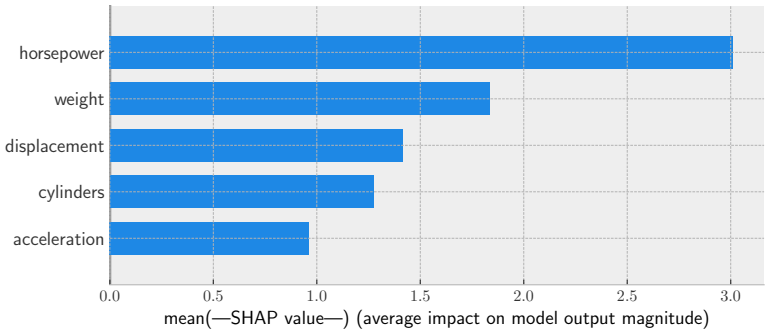
AME-ITEA



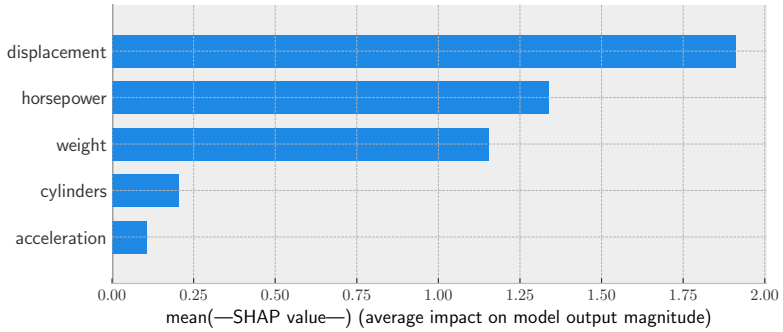
FScore-XGB



SHAP-ITEA

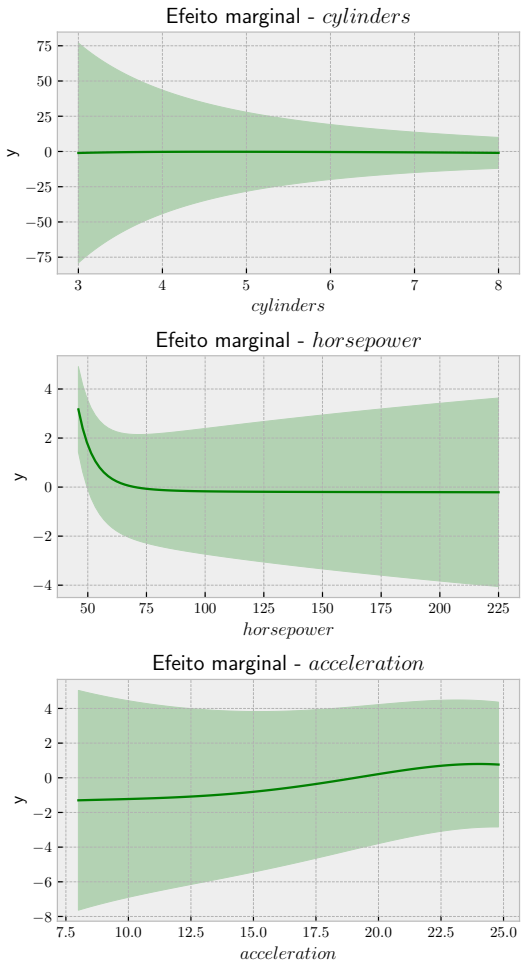


SHAP-XGB

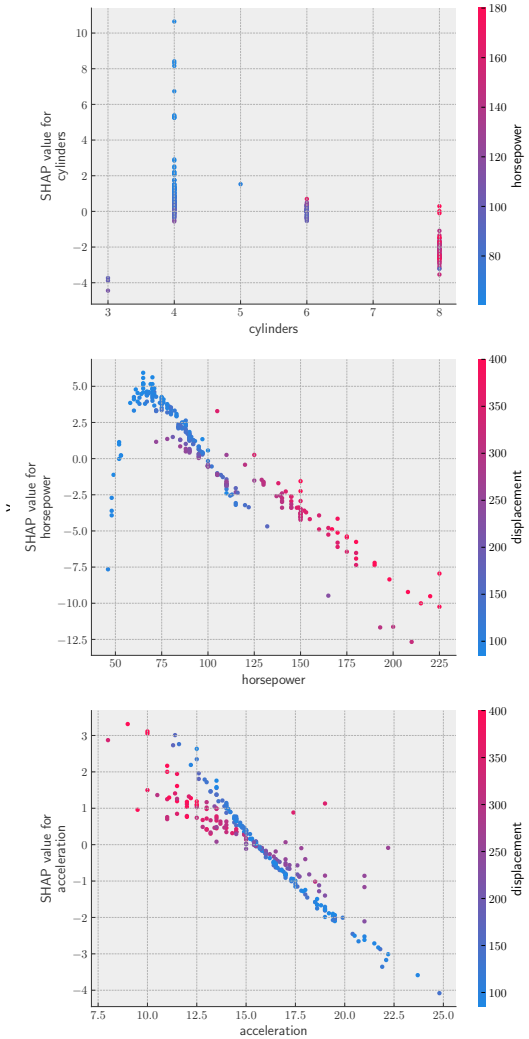


Cars Dataset - Contribuição das variáveis

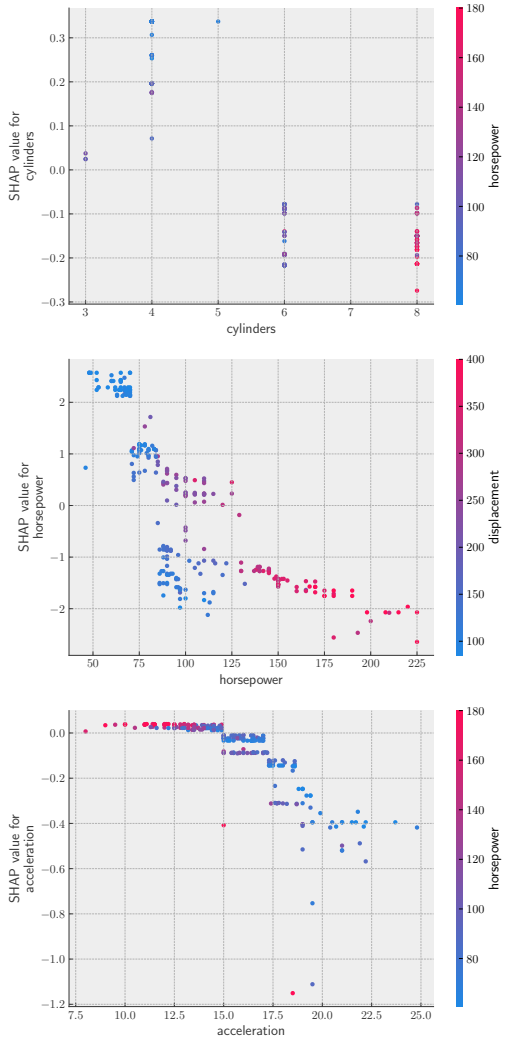
MEM - ITEA



SHAP-ITEA

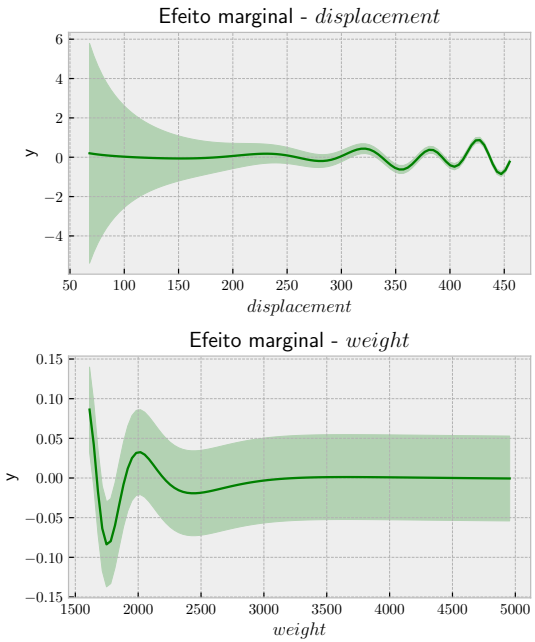


SHAP-XGB

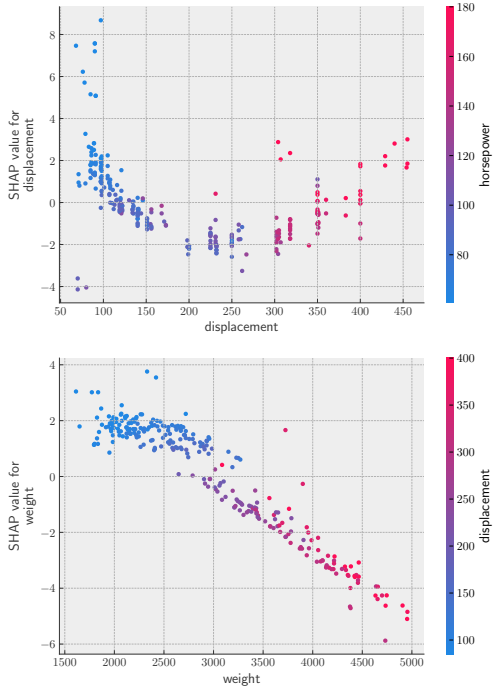


Cars Dataset - Contribuição das variáveis

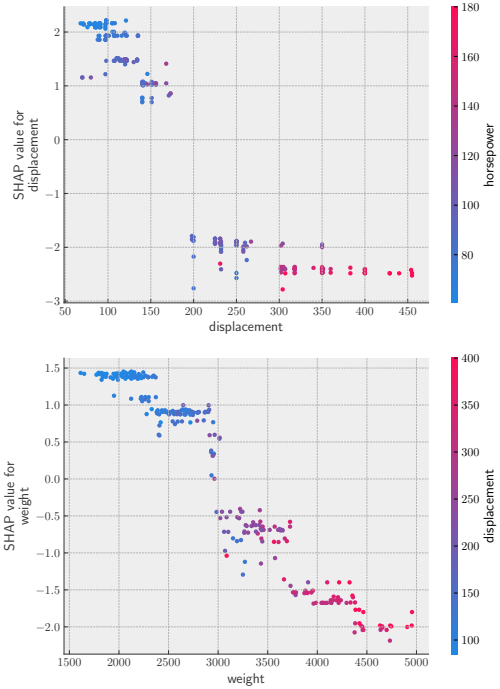
MEM-ITEA



SHAP-ITEA



SHAP-XGB



Índice

Considerações dos experimentos

- *Sanity Check*:
  - mostrou funcionamento correto
  - diferenças entre AME e SHAP
- *Cars Dataset*:
  - Colinearidade entre as variáveis é grande
  - Explicações variam com o método de importância/regressor

Contato

Obrigado!

Guilherme Aldeia

✉ guilherme.aldeia@ufabc.edu.br

Fabrício Olivetti de França

✉ folivetti@ufabc.edu.br

Link da apresentação

galdeia.github.io/presentations/webnarHAL.pdf