



JÁ VAI COMEÇAR

VIÉS, VARIÂNCIA E REGULARIZAÇÃO

ENQUANTO ISSO...

- Escolha um lugar confortável para você sentar e se acomodar
- Que tal pegar um snack pra matar a fome, uma água, um chá
- Abra o chat, envie um "olá" e #sentimento de como chega
- Que tal pegar caderno e caneta para anotações

T

**QUE BOM
QUE VOCÊ
VEIO!**





FELIPE FÁBIO FRIGERI

Staff Data Scientist – Nubank

[LinkedIn](#)

Currículo rápido:

Graduação Física/USP (2010) (+ professor de física por 7 anos)

Mestrado/USP – Física Atmosférica (2013)

Itaú Unibanco – 2013-2021

Ex-aluno Tera! (Turma de Outubro/2019)

VIÉS, VARIÂNCIA E REGULARIZAÇÃO

A vertical bar with a gradient from light green at the top to light blue at the bottom.

ALGUMA EXPECTATIVA
ESPECÍFICA SOBRE OS
TEMAS?

AGENDA

- **Modelos preditivos e generalização**
+ Intervalo – 10 min
- **Entendendo viés e variabilidade na prática**
- **Seleção de variáveis: Conceitos e motivos**
- **Selecionando variáveis na prática**



UM OLHAR MAIS APROFUNDADO SOBRE MODELOS PREDITIVOS

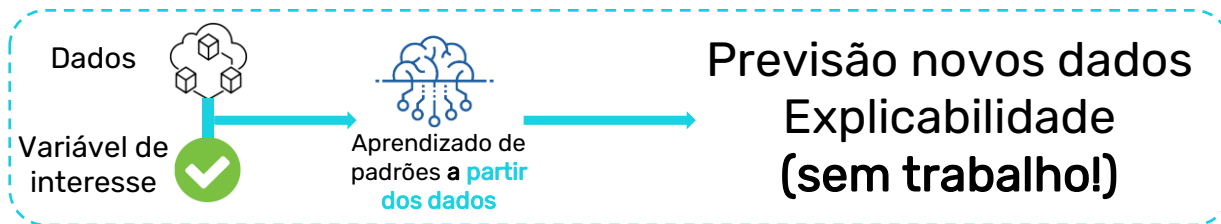
Um pouco sobre o que já sabemos

No **aprendizado de máquina**, partimos de **dados** e uma **variável resposta** para encontrar relações e permitir a realização de **previsões**

Programação Explícita



Aprendizado de máquina

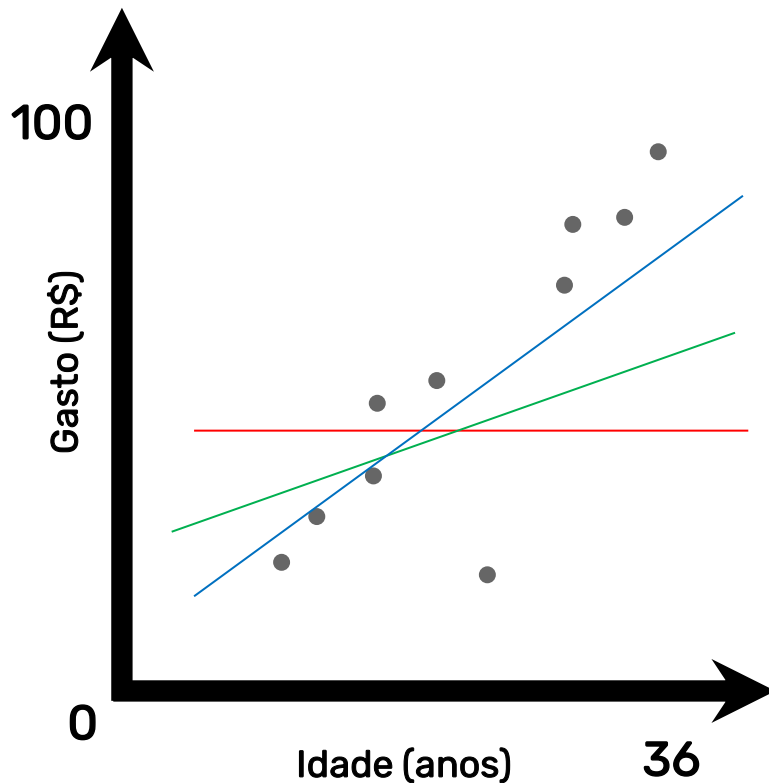


T

Um pouco sobre o que já sabemos

No **aprendizado de máquina**, partimos de **dados** e uma **variável resposta** para encontrar relações e permitir a realização de **previsões**

Id usuário	Idade (anos)	Gasto (R\$)
1	20	40
2	22	42
3	24	46
4	26	52
5	28	56
6	30	32
7	32	50
8	34	50





E porque um modelo preditivo?

Falamos bastante da capacidade do computador **aprender modelos preditivos**. Mas qual é, de fato, a relevância disso?



Tomar decisões inteligentes
antes que o futuro chegue!

Ex: Desconto em produtos para usuários de e-commerce que tendem a gastar mais



**USO
PREDITIVO**



Também podemos aprender
quais dados tem mais relação
com nossa variável de interesse

Ex: O que influencia mais o gasto de um usuário: idade ou época do ano?

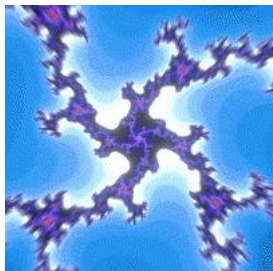


**USO
INFERENCIAL**

T

Se usamos um modelo para prever, o que é realmente importante?

Para que um modelo preditivo seja útil, precisamos que ele tenha algumas características **essenciais**:



Capturar a **complexidade** entre os dados e a variável resposta



As regras e relações permitirem a **previsão** **correta em novos dados**

A vertical bar with a gradient from light green at the top to light blue at the bottom.

E COMO GARANTIR
PREVISÕES CORRETAS?
PODEMOS MEDIR ISSO?



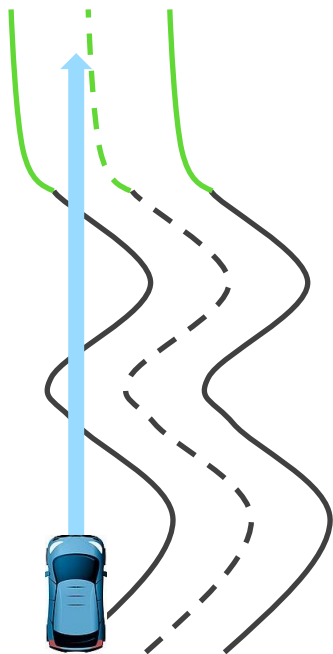
Um exemplo prático: Dirigindo em uma pista

Vamos imaginar uma situação em que motoristas querem **aprender a dirigir em uma estrada**

— Pista de aprendizado — Pista de teste

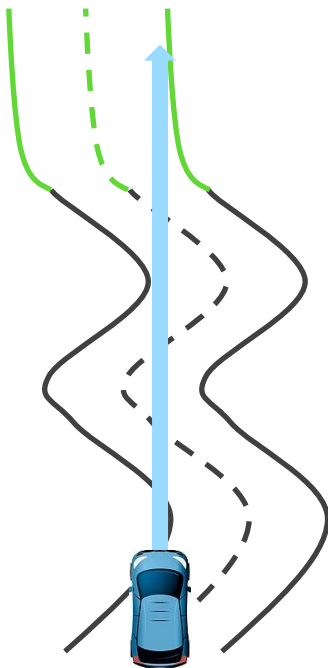
Motorista 1

Linha reta, levemente à esquerda do centro das pistas



Motorista 2

Linha reta, no centro das pistas



Qual dos dois motoristas dirigiu melhor em cada uma das pistas?

	Motorista 1	Motorista 2
Pista de Aprendizado		
Pista de Teste		

Neste caso, o **motorista 1** está cometendo um erro **sistemático** (sempre da mesma forma)

VIÉS (OU BIAS)



Um exemplo prático: Dirigindo em uma pista

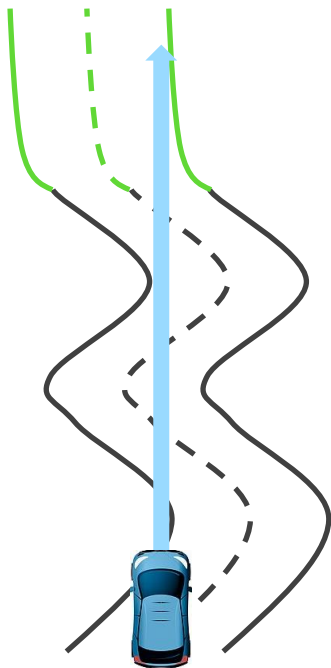
Vamos imaginar uma situação em que motoristas querem **aprender a dirigir em uma estrada**

— Pista de aprendizado

— Pista de teste

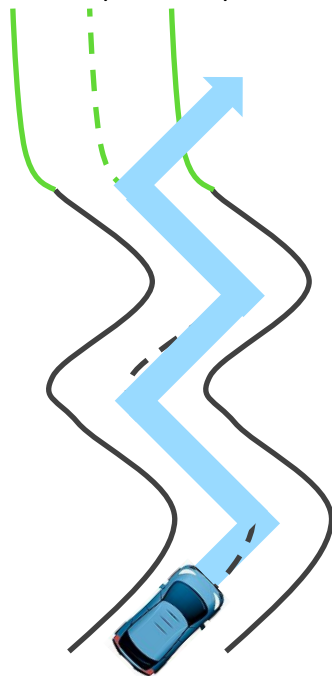
Motorista 2

Linha reta, no centro das pistas



Motorista 3

Faz curvas alternadas de 45 graus, como na pista de aprendizado



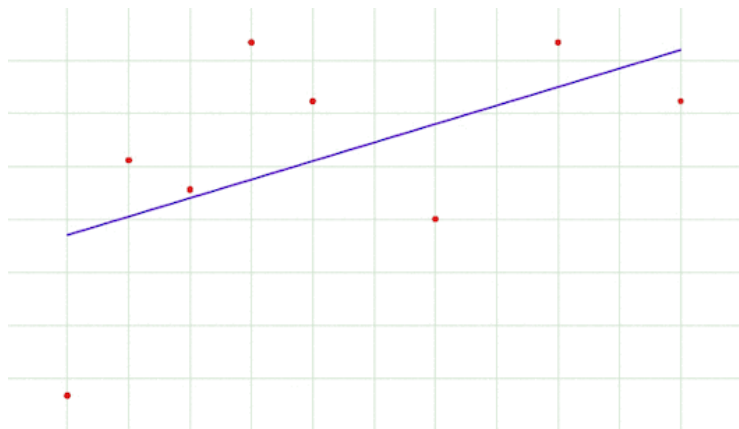
Qual dos dois motoristas dirigiu melhor em cada uma das pistas?

	Motorista 2	Motorista 3
Pista de Aprendizado		
Pista de Teste		

Neste caso, o **motorista 3** decorou a pista de aprendizado, mas isso não foi útil para dirigir na pista de teste, cometendo muitos erros neste trecho

**VARIABILIDADE
(OU VARIANCE)**

T

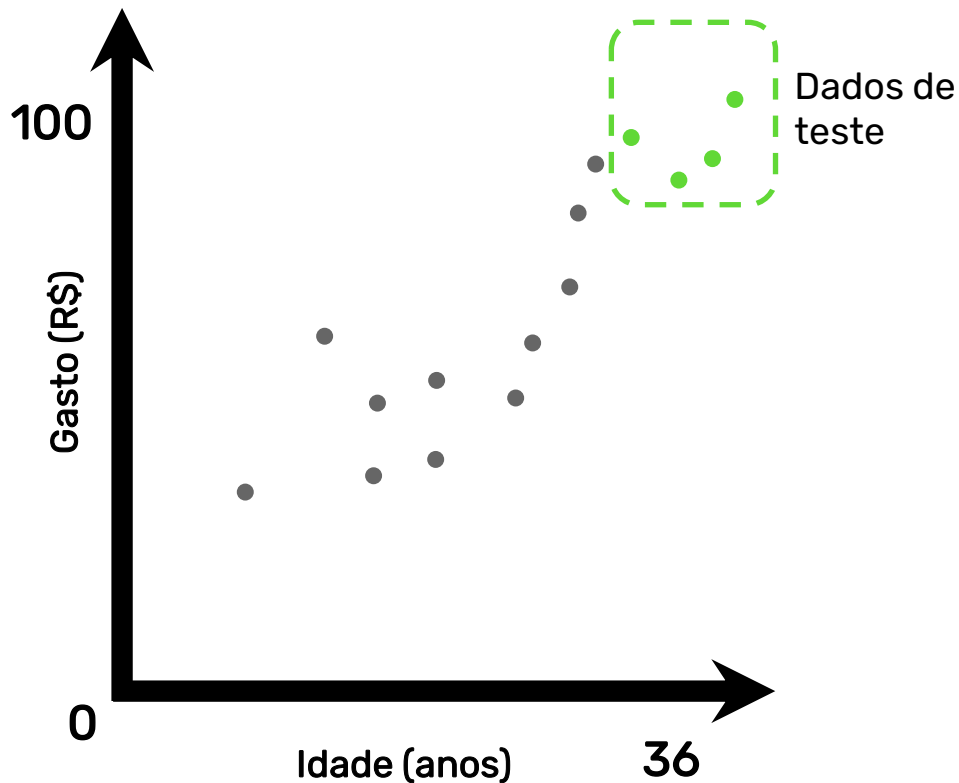


MAS COMO ISSO SE
RELACIONA COM
MODELOS LINEARES?



A relação entre complexidade e acerto

A complexidade dos modelos determina **seu poder de previsão** em dados conhecidos e **novos**

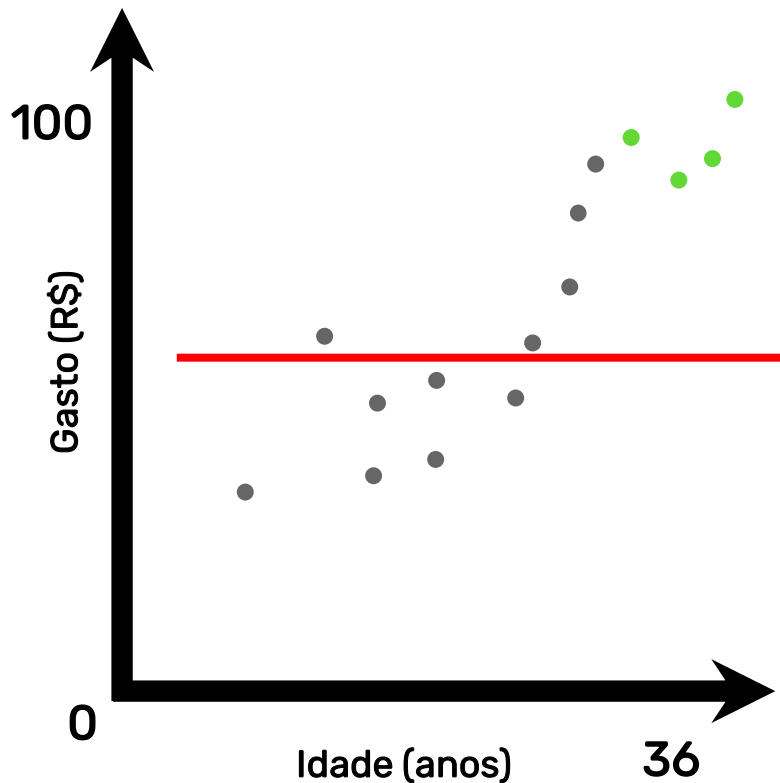






A relação entre complexidade e acerto

A complexidade dos modelos determina **seu poder de previsão** em dados conhecidos e **novos**

Modelo 1: $y = A$ (constante)



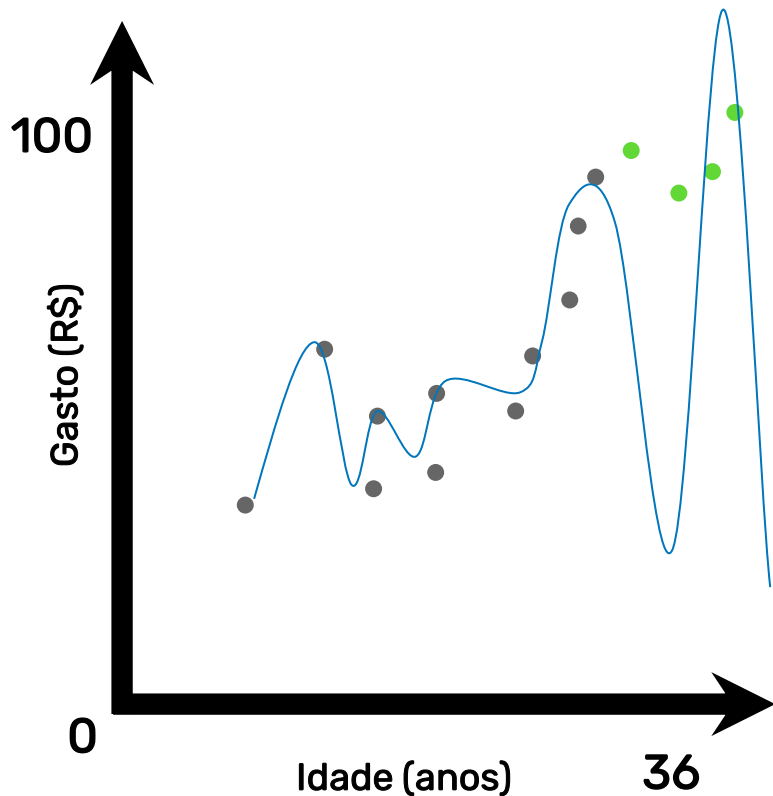
Fórmula	Acerto em dados conhecidos	Acerto em dados novos
$y = A$		







A relação entre complexidade e acerto

A complexidade dos modelos determina **seu poder de previsão** em dados conhecidos e **novos**

Modelo 2: $y = \text{polinômio complexo}$



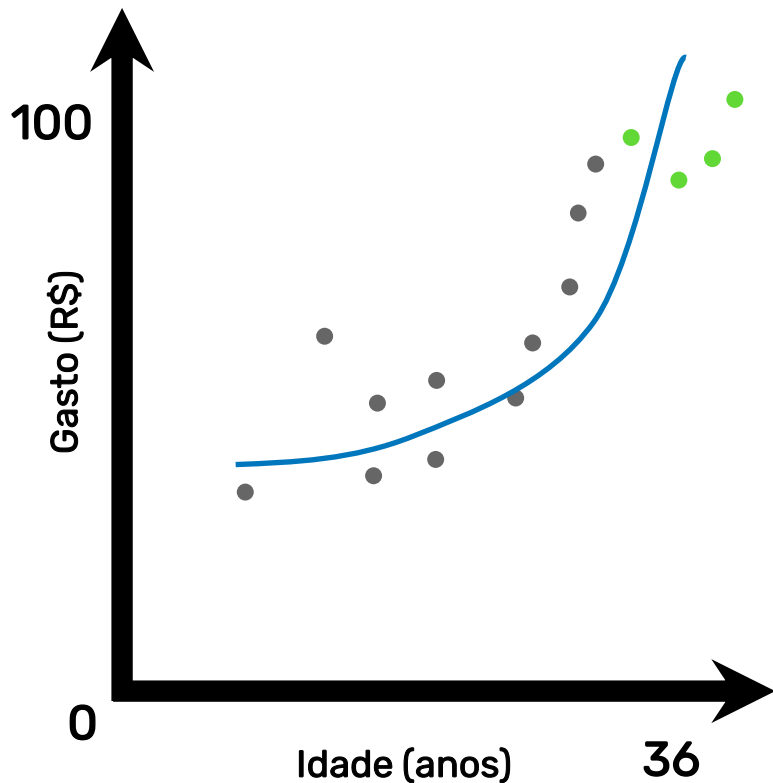
Fórmula	Acerto em dados conhecidos	Acerto em dados novos
$y = A$		
$y = \text{polinômio complexo}$		



A relação entre complexidade e acerto

A complexidade dos modelos determina **seu poder de previsão** em dados conhecidos e **novos**

Modelo 2: $y = A + Bx + Cx^2$

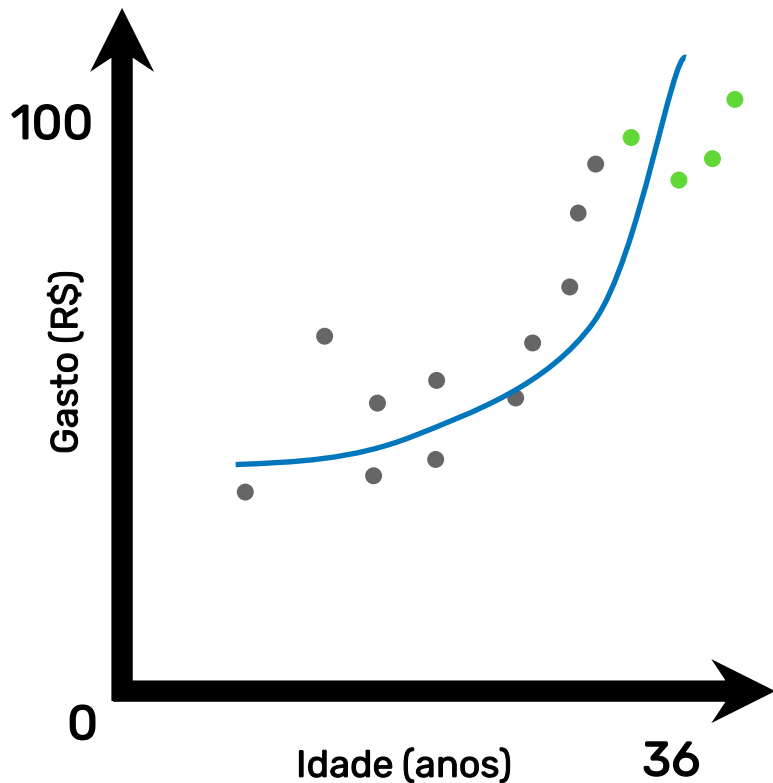


Fórmula	Acerto em dados conhecidos	Acerto em dados novos
$y = A$		
$y = \text{polinômio complexo}$		
$y = A + Bx + Cx^2$		

T A relação entre complexidade e acerto

A complexidade dos modelos determina **seu poder de previsão** em dados conhecidos e **novos**

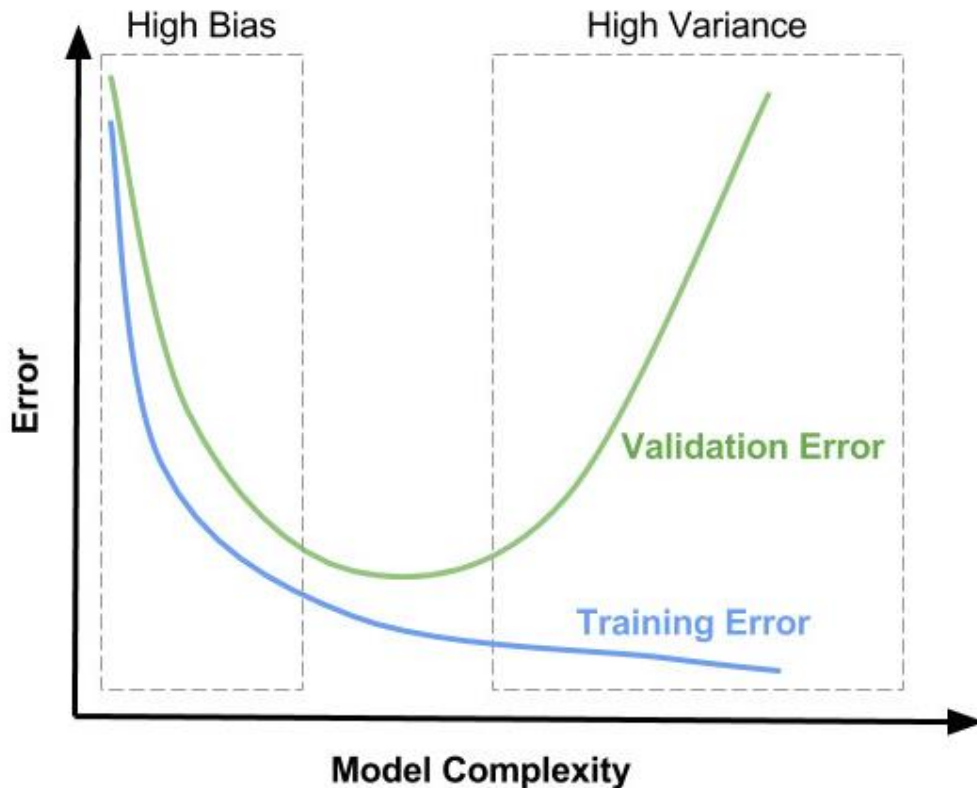
Modelo 2: $y = A + Bx + Cx^2$



Fórmula	Acerto em dados conhecidos	Acerto em dados novos	Cenário
$y = A$			UNDERFITTING (ou subajuste)
$y = \text{polinômio complexo}$			OVERFITTING (ou sobajuste)
$y = A + Bx + Cx^2$			SITUAÇÃO IDEAL

T Tradeoff viés-variabilidade

Como analisamos anteriormente, **complexidade e acerto** possuem uma relação inversa!



T INTERVALO 10 MIN



APROVEITE PARA:


- Fazer anotações do que viu até agora (aprendizados, insights, dúvidas)
- Levantar-se, esticar os braços e as pernas, relaxar por mais tempo
- Comer algo para voltar com energia renovada
- Ir ao banheiro

NO RETORNO, TEREMOS!:

- A aplicação prática de viés e variância no Python, bem como a introdução à **Regularização!**

A vertical bar with a gradient from light green at the top to light blue at the bottom.

IDENTIFICANDO VIÉS E VARIABILIDADE NA PRÁTICA!

A vertical bar with a gradient from light green at the top to light blue at the bottom.

**E QUANDO TEMOS
RECURSOS LIMITADOS?**



Um cenário de decisões difíceis

Não é incomum encontrarmos cenários em que **não conseguimos usar todos os dados disponíveis**, seja por limitação de negócio ou computacional. Como decidir, então, as variáveis mais importantes para nosso projeto?

Exemplo:

Propensão a contratação de produtos

Variáveis disponíveis



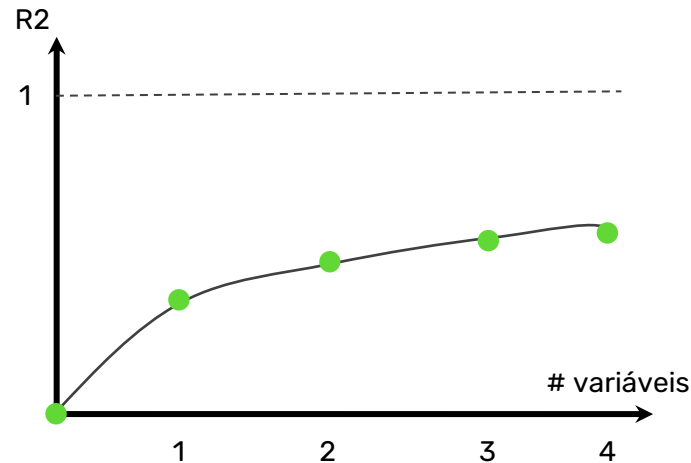
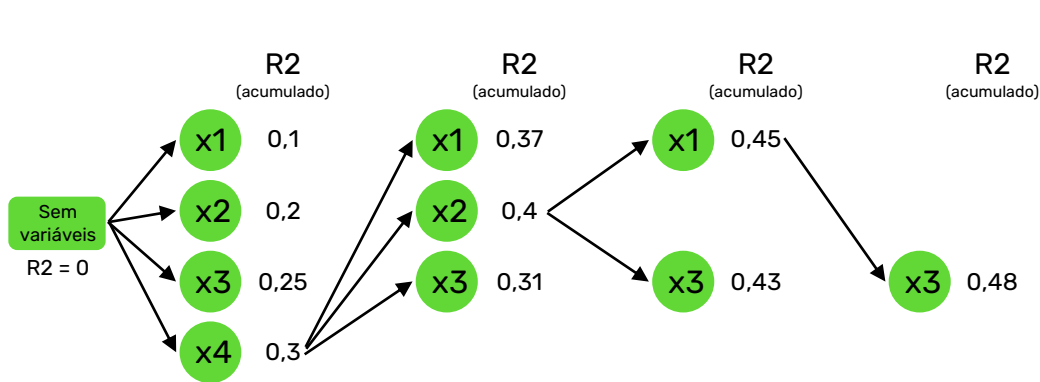
Situações possíveis:

- O sistema de implantação do modelo **não possui alguma variável**
- A operação do negócio **não tem acesso a alguma variável**
- Os servidores da empresa **não suportam o uso de modelos pesados** na velocidade exigida pelo negócio



Selecionando variáveis eficientemente

No **forward feature selection**, partimos de um modelo sem variáveis e evoluímos até o modelo com todas as variáveis, acompanhando a performance a cada nova variável incluída



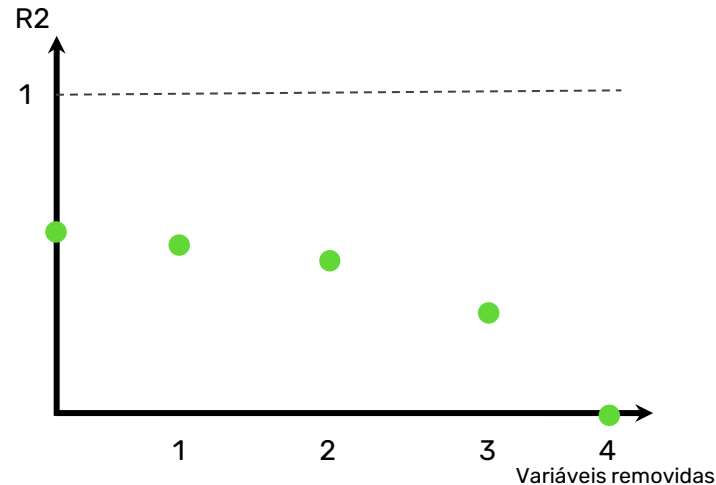
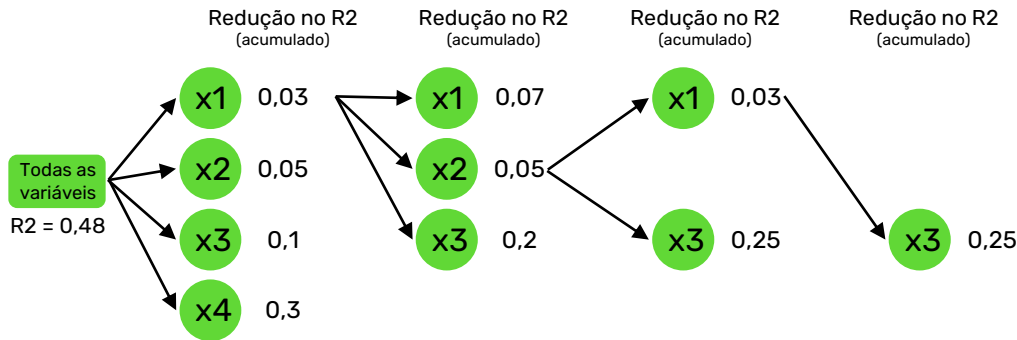
Etapas

- Partimos de um modelo sem variáveis
- **Incorporamos no modelo a variável que mais aumentará o R^2 (ou a métrica da sua preferência)**
- Repetimos o processo até que se atinja o a performance desejada



Selecionando variáveis eficientemente

No **backward feature selection**, partimos de um modelo com todas as variáveis e reduzimos uma a uma, sempre removendo a que menos contribui para a performance do modelo



Etapas

- Partimos de um modelo com todas as variáveis
- **Removevamos do modelo a variável que menos reduzirá o R^2 (ou a métrica da sua preferência)**
- Repetimos o processo até que se atinja o a performance desejada




SELECIONANDO VARIÁVEIS NA PRÁTICA!

T



DÚVIDAS FINAIS

A vertical bar with a gradient from green at the top to blue at the bottom.

COMO FOI?

