

Introdução

Peça para uma IA prever o preço de um imóvel, a partir de um dataset qualquer, e ela devolve um número em segundos. Regressão linear, árvore de decisão, Random Forest, XGBoost, qualquer um desses algoritmos roda sem esforço, produz uma previsão, e até relata uma métrica de erro, se você pedir.

Esse ebook não é sobre isso.

O objetivo aqui não é prever preço de imóvel. É entender, com rigor matemático, por que diferentes algoritmos, aplicados ao mesmo conjunto de dados, chegam a previsões diferentes, e o que essa diferença revela sobre a estrutura do problema, não só sobre qual modelo "ganhou" numa competição de métrica.

Este livro é para cientistas de dados intermediários/avançados que sabem rodar um modelo, mas não sabem por que ele quebra. Ele não ensina a rodar Scikit-learn, ensina porque o modelo que você já roda às vezes quebra, e o que fazer quando isso acontece.

A diferença entre prever e entender

A maioria dos tutoriais de machine learning segue um roteiro parecido. Carregar dados, escolher um algoritmo, treinar, medir erro, ajustar hiperparâmetro, repetir até o número melhorar. Esse processo funciona, no sentido de que produz um modelo que roda. Mas ele raramente responde à pergunta mais interessante, que é **porque aquele algoritmo específico se saiu melhor, ou pior, do que outro, para aquele conjunto de dados específico**.

A resposta para essa pergunta nunca está só no número final de erro. Está na formulação matemática de cada algoritmo, no tipo de relação que cada um consegue capturar, e nas suposições que cada um faz, implicitamente, sobre como os dados se comportam.

O que este ebook vai fazer, de forma diferente

Vamos usar um único [dataset](#), real, imóveis vendidos no condado de Ames, Iowa, e aplicar cinco algoritmos diferentes de regressão sobre ele. Mas em vez de correr direto para a comparação de métrica, vamos primeiro entender a matemática de cada algoritmo, o que ele está de fato calculando por trás, e só depois comparar os resultados, já sabendo o que esperar de cada um, e por quê.

Isso significa que, antes de qualquer previsão, vamos analisar os próprios dados com cuidado. Qual é o formato da distribuição do preço. Quais variáveis realmente se relacionam com ele, e de que forma, linear ou não. Onde existe ausência de dado, e por que ela existe. Onde a variância do erro provavelmente não vai ser constante, antes mesmo de treinar qualquer modelo.

Um exemplo real, que vamos desenvolver já aparece aqui na introdução. Ao explorar o dataset de Ames, encontramos que o desvio-padrão do preço varia enormemente entre bairros, de pouco mais de nove mil dólares num bairro, a mais de cem mil em outro. Essa é uma informação que nenhum algoritmo descobre sozinho, e nenhuma métrica agregada revela, se você não souber que precisa procurar por ela.

Um algoritmo pode reportar erro médio baixo, e ainda assim estar sistematicamente errado num subgrupo específico dos dados, o tipo de problema que discuti extensamente numa série de artigos sobre validação matemática de modelos de IA. A diferença entre um Cientista de Dados que só roda o algoritmo e observa o número final, e um que entende por que aquele número apareceu, está justamente nesse tipo de análise prévia, que a maioria pula.

O escopo deste livro

Vamos trabalhar com cinco algoritmos de regressão, cada um com formulação matemática própria, apresentada em detalhe, sem depender de código para entender o conceito. O código completo, em Python, fica disponível à parte, no [GitHub](#), para quem quiser reproduzir ou estender a análise. Aqui, o foco é o raciocínio matemático por trás de cada configuração, e as consequências de escolher uma abordagem em vez de outra, não a sintaxe de nenhuma biblioteca específica.

Ao final, você não vai só saber qual algoritmo teve o menor erro num dataset específico. Vai entender por que, o que isso revela sobre a estrutura dos dados, e como aplicar esse mesmo raciocínio a qualquer outro problema de regressão que encontrar depois.

Sumário

Introdução

- A diferença entre prever e entender
- O que este ebook vai fazer, de forma diferente
- O escopo deste livro

Capítulo 1 — O Que os Dados Já Revelam, Antes de Qualquer Modelo

- A estrutura básica
- A forma da distribuição do preço
 - Momentos
- Quais variáveis realmente se relacionam com o preço
- Abrindo "Qualidade geral", nível por nível
- O desvio-padrão crescente
- A granularidade da escala, dez níveis, testada diretamente
- A conclusão que une as duas observações
- O achado que vai atravessar o livro inteiro
- O que fazer a seguir
- O que o Capítulo 1 ensina, para o Cientista de Dados

Capítulo 2 — Valores Ausentes, e a Redução para 15 Variáveis

- Revisitando os três mecanismos de ausência
- O que o dataset de Ames revela, olhando de novo
- Onde de fato existe ausência mais próxima de MAR clássico
- Por que não é tão simples quanto “só descartar a coluna”
- O que o Capítulo 2 ensina, para o Cientista de Dados

Capítulo 3 — As Quinze Variáveis Escolhidas, e Por Quê

- Quando "ordem certa" não significa "distância uniforme"
- O que esse padrão repetido confirma
- O que o Capítulo 3 ensina, para o Cientista de Dados

Capítulo 4 — A Base da Comparação: Como Particionar os Dados

- Em que escala modelar o preço?
- Treino, validação e teste, não só treino e teste
- "Cross-Validation", uma alternativa à validação única
- "Bagging", o princípio por trás de "Random Forest"

"Boosting", o princípio por trás de "Gradient Boosting"

O problema de medir erro só no treino

O mesmo cuidado vale para qualquer transformação, não só para a avaliação

O que o Capítulo 4 ensina, para o Cientista de Dados

Capítulo 5 — Mínimos Quadrados, Regularização, OLS e Gradient Descent

Relembrando a matemática

Forma matricial

O Número de Condição

Exemplo Numérico

Primeiro, sem multicolinearidade

Adicionando uma terceira variável quase idêntica

As suposições implícitas, e o que acontece quando falham

Regularização

"Ridge", "Lasso" e "Elastic Net"

O que este Capítulo 5 ensina, para o Cientista de Dados

Capítulo 6 — Regressão Linear Aplicada: Multicolinearidade e Regularização no Dataset de Ames

Dois casos reais, com intensidade diferente

O que acontece com os coeficientes

Multicolinearidade, Via VIF o que é?

O que o Capítulo 6 ensina, para o Cientista de Dados

Capítulo 7 — Árvore de Decisão para Regressão: Reduzindo Variância, Não Entropia

O mesmo princípio, medido de outro jeito

Importância

A fórmula da importância total do modelo

O que a árvore prevê, em cada folha

Capítulo 8 — Árvore de Decisão Aplicada ao Dataset de Ames

Revisitando o overfitting do Capítulo 3

A primeira divisão da árvore, com dado real

Como a árvore lida com a variável "Bairro", e o problema que isso revela

"Target Encoding", a correção

Regularização na Árvore de Decisão

O Pruning (Poda) como Regularização Explícita

Outros Hiperparâmetros como Mecanismos de Regularização (Pre-pruning)

Onde árvore isolada ainda vence, na prática, mesmo perdendo em RMSE

O que os Capítulos 7 e 8 ensinam, juntos, para o Cientista de Dados

Capítulo 9 — Random Forest: Muitas Árvores, Menos Variância

"Bagging", formalizado

A redução da variância

O segundo ingrediente, que "Random Forest" completo adiciona

O que o Capítulo 9 ensina para o Cientista de Dados

Capítulo 10 — Random Forest Aplicado ao Dataset de Ames

Confirmando a fórmula do Capítulo 9, com número real

O que aconteceu com "Bairro", a mesma fragmentação de antes

"Out-of-Bag Score", uma validação embutida

O que o Capítulo 10 ensina para o Cientista de Dados

Capítulo 11 — Gradient Boosting: Corrigindo Erros, Um Passo de Cada Vez

Uma lógica completamente diferente de "Bagging"

A construção matemática do modelo

Por que "gradiente", especificamente

"Weak learners"

Por que usar um modelo ruim, de propósito, funciona tão bem

Retomando o dataset de cinco imóveis

O que o Capítulo 11 ensina para o Cientista de Dados

Capítulo 12 — Gradient Boosting Aplicado ao Dataset de Ames

Capítulo 13 — XGBoost: Um Ajuste Mais Fino do Ajuste

Capítulo 14 — XGBoost Aplicado ao Dataset de Ames

A comparação controlada

Por que "XGBoost" não melhorou, neste caso

Testando a hipótese, com o dataset completo

Capítulo 15 — Outliers e Vazamento de Dados: Dois Erros Diferentes, Duas Correções Diferentes

Outliers, revisitados com mais precisão

Vazamento de dados, como detectar de verdade

O que de fato funciona, para detectar

O que o Capítulo 15 ensina para o Cientista de Dados

Capítulo 16 — Feature Engineering: Quando Combinar Variáveis Vale Mais Que Cada Uma Isolada

Dois tipos de combinação, com resultados opostos

Primeiro exemplo, uma combinação linear, "Idade do imóvel na venda"

Segundo exemplo, uma interação, "Qualidade geral" multiplicada por "Área útil habitável"

Por que a diferença entre os dois exemplos é tão grande

Testando se essa foi, de fato, a melhor escolha possível

O que este capítulo ensina para o Cientista de Dados

Capítulo 17 — Dados Não Estacionários: Quando o Passado Deixa de Prever o Futuro

O teste

O que essa comparação revela

Por que "Random Forest" é o mais sensível à mudança de regime

O que o Capítulo 17 ensina para o Cientista de Dados

Capítulo 18 — Heterocedasticidade Revisitada: Diagnóstico e Correção com Dado Real

O diagnóstico, com número real, não mais simulado

Breusch-Pagan, o teste de diagnóstico

Mínimos Quadrados Ponderados, "WLS", a correção

O que o Capítulo 18 ensina para o Cientista de Dados

Capítulo 19 — A Comparação Final: Por Que os Cinco Algoritmos Discordam

O que a comparação com material público sobre este mesmo dataset revela

Uma divergência real, e reveladora

Os cinco, lado a lado, sobre o mesmo dado

O que este livro ensinou, além dos números de um dataset específico

Quando faz sentido trabalhar com árvore isolada

Quando faz sentido trabalhar com Regressão Linear pura

A pergunta que você deveria carregar para o próximo problema