
Projetando Sistemas de Machine Learning

*Processo Iterativo para
Aplicações Prontas para Produção*

Chip Huyen



ALTA BOOKS

GRUPO EDITORIAL

Rio de Janeiro, 2023

Projetando Sistemas de Machine Learning

Copyright © 2023 ALTA BOOKS

ALTA BOOKS é uma empresa do Grupo Editorial Alta Books (Starlin Alta Editora e Consultoria Ltda.)

Copyright © 2022 Huyen Thi Khanh Nguyen.

ISBN: 978-85-508-1967-9

Authorized Portuguese translation of the English edition of *Designing Machine Learning Systems* ISBN 9781098107963 © 2022 Huyen Thi Khanh Nguyen. This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

PORTUGUESE language edition published by Grupo Editorial Alta Books Ltda., Copyright © 2023 by STARLIN ALTA EDITORA E CONSULTORIA LTDA.

Impresso no Brasil – 1ª Edição, 2023 – Edição revisada conforme o Acordo Ortográfico da Língua Portuguesa de 2009.

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

H987p	Huyen, Chip
Projetando Sistemas de Machine Learning: Processo Iterativo para Aplicações Prontas para Produção / Chip Huyen ; traduzido por Cibelle Ravaglia. - Rio de Janeiro : Alta Books, 2023. 384 p. ; 15,8cm x 23cm.	
Tradução de: <i>Designing Machine Learning Systems</i> . Inclui índice. ISBN: 978-85-508-1967-9	
1. Ciências da computação. 2. Tecnologia. I. Ravaglia, Cibelle. II. Título.	
2023-1218	CDD 004 CDU 004

Elaborado por Wagner Rodolfo da Silva - CRB-8/9410

Índice para catálogo sistemático:

1. Ciências da computação 004
2. Ciências da computação 004

Todos os direitos estão reservados e protegidos por Lei. Nenhuma parte deste livro, sem autorização prévia por escrito da editora, poderá ser reproduzida ou transmitida. A violação dos Direitos Autorais é crime estabelecido na Lei nº 9.610/98 e com punição de acordo com o artigo 184 do Código Penal.

O conteúdo desta obra foi formulado exclusivamente pelo(s) autor(es).

Marcas Registradas: Todos os termos mencionados e reconhecidos como Marca Registrada e/ou Comercial são de responsabilidade de seus proprietários. A editora informa não estar associada a nenhum produto e/ou fornecedor apresentado no livro.

Material de apoio e erratas: Se parte integrante da obra e/ou por real necessidade, no site da editora o leitor encontrará os materiais de apoio (download), errata e/ou quaisquer outros conteúdos aplicáveis à obra. Acesse o site www.altabooks.com.br e procure pelo título do livro desejado para ter acesso ao conteúdo..

Suporte Técnico: A obra é comercializada na forma em que está, sem direito a suporte técnico ou orientação pessoal/exclusiva ao leitor.

A editora não se responsabiliza pela manutenção, atualização e idioma dos sites, programas, materiais complementares ou similares referidos pelos autores nesta obra.

Produção Editorial: Grupo Editorial Alta Books

Diretor Editorial: Anderson Vieira

Vendas Governamentais: Cristiane Mutús

Gerência Comercial: Claudio Lima

Gerência Marketing: Andréa Guatiello

Assistente Editorial: Brenda Rodrigues

Tradução: Cibelle Ravaglia

Copidesque: Rafael Surgek

Revisão: Vinicius Barreto; Thamiris Leiroza

Diagramação: Joyce Matos

Revisão Técnica: Ismael Soares

(Especialista em Machine Learning)



Rua Viúva Cláudio, 291 – Bairro Industrial do Jacaré
CEP: 20.970-031 – Rio de Janeiro (RJ)
Tels.: (21) 3278-8069 / 3278-8419
www.altabooks.com.br – altabooks@altabooks.com.br
Ouvidoria: ouvidoria@altabooks.com.br



Sumário

Prefácio.....	vii
1. Visão Geral dos Sistemas de Machine Learning.....	1
Quando Usar o Machine Learning	3
Entendendo os Sistemas de Machine Learning	13
Recapitulando	25
2. Introdução ao Design de Sistemas de Machine Learning.....	27
Objetivos de Negócios e de ML	28
Requisitos para os Sistemas de ML	31
Processo Iterativo	34
Delimitando os Problemas de ML	37
Mente versus Dados (Data Overmind)	45
Recapitulando	48
3. Fundamentos de Engenharia de Dados	51
Fontes de Dados	52
Formatos de Dados	55
Modelos de Dados	60
Mecanismos de Armazenamento de Dados e Processamento	70
Modos de Dataflow	75
Processamento em Lote versus Processamento de Fluxo	80
Recapitulando	82
4. Treinando os Dados.....	85
Amostragem	86
Rotulagem	92
Classes Desbalanceadas	107
Data Augmentation	118
Recapitulando	122
5. Engenharia de Features	123
Features Aprendidas versus Features Projetadas	123

Operações Comuns de Engenharia de Features	126
Data Leakage	138
Engenharia de Boas Features	144
Recapitulando	148
6. Desenvolvimento de Modelo e Avaliação Offline	151
Desenvolvimento e Treinamento de Modelos	152
Avaliação Offline do Modelo	181
Recapitulando	191
7. Serviço de Predição e Deploy de Modelo	193
Mitos sobre o Deploy de Machine Learning	196
Predição em Lote versus Predição Online	199
Compressão de Modelo	208
ML na Nuvem e na Borda	214
Recapitulando	224
8. Mudanças na Distribuição de Dados e Monitoramento	227
Causas de Falhas de Um Sistema de ML	228
Mudanças na Distribuição de Dados	238
Monitoramento e Observabilidade	252
Recapitulando	263
9. Aprendizado Contínuo e Teste em Produção	265
Aprendizado Contínuo	266
Teste em Produção	284
Recapitulando	294
10. Infraestrutura e Ferramentas para MLOps	297
Armazenamento e Processamento Computacional	301
Ambiente de Desenvolvimento	307
Gerenciamento de Recursos	315
Plataforma de ML	324
Construir versus Comprar	332
Recapitulando	334
11. O Lado Humano do Machine Learning	337
Experiência do Usuário	337
Estrutura de Equipe	341
IA Responsável	345
Recapitulando	359
Epílogo	361
Índice	367