

# CONSTRUA UMA CARREIRA EM Ciência de Dados

Emily Robinson  
Jacqueline Nolis



 MANNING

novatec

# **Construa uma Carreira em Ciência de Dados**

**Emily Robinson  
Jacqueline Nolis**

**MANNING**  
Novatec

Original English language edition published by Manning Publications Co., Copyright © 2020 by Manning Publications. Portuguese-language edition for Brazil copyright © 2021 by Novatec Editora. All rights reserved.

Edição original em Inglês publicada pela Manning Publications Co., Copyright © 2020 pela Manning Publications. Edição em Português para o Brasil copyright © 2021 pela Novatec Editora. Todos os direitos reservados.

© Novatec Editora Ltda. [2021].

Todos os direitos reservados e protegidos pela Lei 9.610 de 19/02/1998. É proibida a reprodução desta obra, mesmo parcial, por qualquer processo, sem prévia autorização, por escrito, do autor e da Editora.

Editor: Rubens Prates

Tradução: Francine Facchin Esteves

Revisão gramatical: Maria Rita Quintella

ISBN do ebook: 978-65-86057-43-0

ISBN do impresso: 978-65-86057-42-3

Histórico de impressões:

Janeiro/2021 Primeira edição

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110

02460-000 – São Paulo, SP – Brasil

Tel.: +55 11 2959-6529

Email: [novatec@novatec.com.br](mailto:novatec@novatec.com.br)

Site: <https://novatec.com.br>

Twitter: [twitter.com/novateceditora](https://twitter.com/novateceditora)

Facebook: [facebook.com/novatec](https://facebook.com/novatec)

LinkedIn: [linkedin.com/in/novatec](https://linkedin.com/in/novatec)

GRA20210107

*De Emily para Michael*  
*e*  
*de Jacqueline para Heather, Amber e Laura,*  
*pelo amor e apoio dados nesta jornada.*

# Sumário

## [Prefácio](#)

## [Agradecimentos](#)

## [Sobre este livro](#)

## [Sobre as autoras](#)

## [Sobre a ilustração da capa](#)

## [parte I ■ Introdução à ciência de dados](#)

### [capítulo 1 ■ O que é a ciência de dados?](#)

#### [1.1 O que é a ciência de dados?](#)

##### [1.1.1 Matemática/estatística](#)

##### [1.1.2 Bancos de dados/programação](#)

##### [1.1.3 Entendimento dos negócios](#)

#### [1.2 Tipos diferentes de vagas na ciência de dados](#)

##### [1.2.1 Análise](#)

##### [1.2.2 Machine learning](#)

##### [1.2.3 Ciência de decisão](#)

##### [1.2.4 Trabalhos relacionados](#)

#### [1.3 Como escolher sua direção](#)

#### [1.4 Entrevista com Robert Chang, cientista de dados da Airbnb](#)

#### [Resumo](#)

### [capítulo 2 ■ Empresas de ciência de dados](#)

#### [2.1 MTC \(Massive Tech Company – Empresa de tecnologia em massa\)](#)

##### [2.1.1 Sua equipe: uma de muitas na MTC](#)

##### [2.1.2 A tecnologia: avançada, mas desarticulada na empresa](#)

##### [2.1.3 As vantagens e desvantagens da MTC](#)

#### [2.2 HandbagLOVE: a varejista bem estabelecida](#)

##### [2.2.1 Sua equipe: um pequeno grupo com dificuldade para crescer](#)

##### [2.2.2 Sua tecnologia: recursos legados que estão começando a mudar](#)

##### [2.2.3 Vantagens e desvantagens da HandbagLOVE](#)

#### [2.3 Seg-Metra: a startup em fase inicial](#)

##### [2.3.1 Sua equipe \(que equipe?\)](#)

##### [2.3.2 A tecnologia: tecnologia de ponta](#)

##### [2.3.3 As vantagens e desvantagens da Seg-Metra](#)

#### [2.4 Videory: a bem-sucedida startup de tecnologia de estágio avançado](#)



- [2.4.1 A equipe: especializada, mas com espaço para mover-se](#)
- [2.4.2 A tecnologia: evitando se complicar com código legado](#)
- [2.4.3 As vantagens e as desvantagens da Videory](#)
- [2.5 Global Aerospace Dynamics: uma gigante do governo](#)
  - [2.5.1 A equipe: um cientista de dados em um mar de engenheiros](#)
  - [2.5.2 A tecnologia: antiga, endurecida e com bloqueio de segurança](#)
  - [2.5.3 As vantagens e desvantagens da GAD](#)
- [2.6 Reunindo tudo](#)
- [2.7 Entrevista com Randy Au, pesquisador quantitativo de experiência do usuário na Google](#)
- [Resumo](#)

## **capítulo 3 ■ Como aprender as competências**

- [3.1 Curso universitário em ciência de dados](#)
  - [3.1.1 Escolha da universidade](#)
  - [3.1.2 Como ingressar em um curso universitário](#)
  - [3.1.3 Resumo de cursos acadêmicos](#)
- [3.2 Participação de bootcamps](#)
  - [3.2.1 O que você aprende](#)
  - [3.2.2 Custo](#)
  - [3.2.3 Escolha do curso](#)
  - [3.2.4 Resumo dos bootcamps de ciência de dados](#)
- [3.3 Trabalhar com ciência de dados dentro de sua empresa](#)
  - [3.3.1 Resumo de aprender no trabalho](#)
- [3.4 Aprender por conta própria](#)
  - [3.4.1 Resumo de aprender por conta própria](#)
- [3.5 Escolha](#)
- [3.6 Entrevista com Julia Silge, cientista de dados e engenheira de software da RStudio](#)
- [Resumo](#)

## **capítulo 4 ■ Como montar um portfólio**

- [4.1 Como criar um projeto](#)
  - [4.1.1 Como encontrar dados e fazer uma pergunta](#)
  - [4.1.2 A escolha de uma direção](#)
  - [4.1.3 Preenchimento do README do GitHub](#)
- [4.2 Como iniciar um blog](#)
  - [4.2.1 Tópicos em potencial](#)
  - [4.2.2 Logística](#)
- [4.3 Trabalhar em projetos exemplares](#)
  - [4.3.1 Freelancers de ciência de dados](#)
  - [4.3.2 Treinar uma rede neural em placas de carro ofensivas](#)
- [4.4 Entrevista com David Robinson, cientista de dados](#)
- [Resumo](#)
- [Recursos dos capítulos 1–4](#)
  - [Livros](#)
  - [Textos de blog](#)

## **parte II ■ Como encontrar empregos em ciência de dados**

### **capítulo 5 ■ A busca: como identificar o emprego certo para você**

#### 5.1 Como encontrar vagas

##### 5.1.1 Como decodificar as descrições

##### 5.1.2 Como detectar sinais de alertas

##### 5.1.3 Definição das suas expectativas

##### 5.1.4 Participação de meetups

##### 5.1.5 Utilização de redes sociais

#### 5.2 Como decidir a quais vagas se candidatar

#### 5.3 Entrevista com Jesse Mostipak, promotora na Kaggle

#### Resumo

### **capítulo 6 ■ Como se candidatar: currículo e carta de apresentação**

#### 6.1 Currículo: o básico

##### 6.1.1 Estrutura

##### 6.1.2 Como aprofundar a seção de experiência: geração de conteúdo

#### 6.2 Carta de apresentação: o básico

##### 6.2.1 Estrutura

#### 6.3 Adaptação

#### 6.4 Indicações

#### 6.5 Entrevista com Kristen Kehrer, instrutora de ciência de dados e criadora de curso

#### Resumo

### **capítulo 7 ■ A entrevista: o que esperar e como lidar com ela**

#### 7.1 O que as empresas querem?

##### 7.1.1 O processo de entrevista

#### 7.2 Etapa 1: a triagem inicial por telefone

#### 7.3 Etapa 2: entrevista na empresa

##### 7.3.1 A entrevista técnica

##### 7.3.2 A entrevista comportamental

#### 7.4 Etapa 3: o estudo de caso

#### 7.5 Etapa 4: a entrevista final

#### 7.6 A proposta

#### 7.7 Entrevista com Ryan Williams, cientista de decisão sênior da Starbucks

#### Resumo

### **capítulo 8 ■ A proposta: saber o que aceitar**

#### 8.1 O processo

#### 8.2 Como receber a proposta

#### 8.3 A negociação

- [8.3.1 O que é negociável?](#)
- [8.3.2 Quanto negociar](#)
- [8.4 Táticas de negociação](#)
- [8.5 Como escolher entre duas “boas” propostas de trabalho](#)
- [8.6 Entrevista com Brooke Watson Madubuonwu, cientista de dados sênior da ACLU](#)
- [Resumo](#)
- [Recursos dos capítulos 5–8](#)
  - [Livros](#)
  - [Posts de blog e cursos](#)

## **parte III ■ Adaptação na área de ciência de dados**

### **capítulo 9 ■ Os primeiros meses de trabalho**

- [9.1 Primeiro mês](#)
  - [9.1.1 Integração em uma grande empresa: uma máquina bem lubrificada](#)
  - [9.1.2 Integração em uma empresa pequena: que integração?](#)
  - [9.1.3 Compreender e definir expectativas](#)
  - [9.1.4 Como conhecer seus dados](#)
- [9.2 Como tornar-se produtivo](#)
  - [9.2.1 Como fazer perguntas](#)
  - [9.2.2 Como construir relacionamentos](#)
- [9.3 Se você for o primeiro cientista de dados](#)
- [9.4 Quando o trabalho não é o que foi prometido](#)
  - [9.4.1 O trabalho é terrível](#)
  - [9.4.2 O ambiente de trabalho é tóxico](#)
  - [9.4.3 Decidir ir embora](#)
- [9.5 Entrevista com Jarvis Miller, cientista de dados do Spotify](#)
- [Resumo](#)

### **capítulo 10 ■ Como fazer uma análise eficaz**

- [10.1 Solicitação](#)
- [10.2 Plano de análise](#)
- [10.3 Como fazer a análise](#)
  - [10.3.1 Importação e limpeza de dados](#)
  - [10.3.2 Exploração e modelagem de dados](#)
  - [10.3.3 Pontos importantes para explorar e modelar](#)
- [10.4 Encerramento](#)
  - [10.4.1 Apresentação final](#)
  - [10.4.2 Como salvar seu trabalho para a posterioridade](#)
- [10.5 Entrevista com Hilary Parker, cientista de dados na Stitch Fix](#)
- [Resumo](#)

### **capítulo 11 ■ Como implantar um modelo na produção**

- [11.1 O que significa implantar na produção?](#)
- [11.2 Como criar o sistema de produção](#)
  - [11.2.1 Coleta de dados](#)

- [11.2.2 Como construir o modelo](#)
- [11.2.3 Como atender modelos com APIs](#)
- [11.2.4 Como construir uma API](#)
- [11.2.5 Documentação](#)
- [11.2.6 Teste](#)
- [11.2.7 Como implantar uma API](#)
- [11.2.8 Teste de carregamento](#)
- [11.3 Como manter o sistema em funcionamento](#)
  - [11.3.1 Monitoramento do sistema](#)
  - [11.3.2 Como retreinar o modelo](#)
  - [11.3.3 Fazer alterações](#)
- [11.4 Encerramento](#)
- [11.5 Entrevista com Heather Nolis, engenheira de machine learning da T-Mobile](#)
- [Resumo](#)

## **capítulo 12 ■ Como trabalhar com stakeholders**

- [12.1 Tipos de stakeholders](#)
  - [12.1.1 Stakeholders da administração](#)
  - [12.1.2 Stakeholders da engenharia](#)
  - [12.1.3 Liderança corporativa](#)
  - [12.1.4 Seu gerente](#)
- [12.2 Como trabalhar com stakeholders](#)
  - [12.2.1 Como entender os objetivos dos stakeholders](#)
  - [12.2.2 Como comunicar constantemente](#)
  - [12.2.3 Como ser consistente](#)
- [12.3 Como priorizar o trabalho](#)
  - [12.3.1 Trabalho inovador e com impacto](#)
  - [12.3.2 Trabalho não inovador, mas com impacto](#)
  - [12.3.3 Trabalho inovador, mas sem impacto](#)
  - [12.3.4 Trabalho nem inovador nem impactante](#)
- [12.4 Observações finais](#)
- [12.5 Entrevista com Sade Snowden-Akintunde, cientista de dados da Etsy](#)
- [Resumo](#)
- [Recursos dos capítulos 9–12](#)
  - [Livros](#)
  - [Blogs](#)

## **parte IV ■ Como crescer em sua função em ciência de dados**

### **capítulo 13 ■ Quando seu projeto de ciência de dados falha**

- [13.1 Por que seu projeto de ciência de dados falha?](#)
  - [13.1.1 Os dados não são aqueles que você queria](#)
  - [13.1.2 Os dados não têm um sinal](#)
  - [13.1.3 O cliente acabou não querendo](#)

[13.2 Gerenciamento de risco](#)

[13.3 O que fazer quando seu projeto falhar](#)

[13.3.1 O que fazer com o projeto](#)

[13.3.2 Lidando com emoções negativas](#)

[13.4 Entrevista com Michelle Keim, chefe da equipe de ciência de dados e machine learning da Pluralsight](#)

[Resumo](#)

## **capítulo 14 ■ Como participar da comunidade de ciência de dados**

[14.1 Como aumentar seu portfólio](#)

[14.1.1 Mais posts de blog](#)

[14.1.2 Mais projetos](#)

[14.2 Participação em conferências](#)

[14.2.1 Como lidar com a ansiedade social](#)

[14.3 Como ministrar palestras](#)

[14.3.1 Como conseguir uma oportunidade](#)

[14.3.2 Preparação](#)

[14.4 Como contribuir para o código aberto](#)

[14.4.1 Como contribuir para o trabalho de outras pessoas](#)

[14.4.2 Como fazer seu próprio pacote ou biblioteca](#)

[14.5 Como reconhecer e evitar a exaustão](#)

[14.6 Entrevista com Renee Teate, diretora de ciência de dados na HelioCampus](#)

[Resumo](#)

## **capítulo 15 ■ Como sair bem do seu emprego**

[15.1 Decidir sair](#)

[15.1.1 Faça o balanço do seu progresso de aprendizagem](#)

[15.1.2 Verifique como você se alinha com seu gerente](#)

[15.2 Como a busca de emprego difere do seu primeiro emprego](#)

[15.2.1 Decidir o que você quer](#)

[15.2.2 Entrevistas](#)

[15.3 Encontrar um novo emprego enquanto estiver empregado](#)

[15.4 Como apresentar o aviso-prévio](#)

[15.4.1 Considerando uma contraproposta](#)

[15.4.2 Como contar à equipe](#)

[15.4.3 Como facilitar a transição](#)

[15.5 Entrevista com Amanda Casari, gerente de engenharia da Google](#)

[Resumo](#)

## **capítulo 16 ■ Como subir na carreira**

[16.1 A via da gerência](#)

[16.1.1 Vantagens de ser gerente](#)

[16.1.2 Desvantagens de ser gerente](#)

[16.1.3 Como se tornar um gerente](#)

[16.2 O caminho para ser cientista de dados líder](#)

[16.2.1 Benefícios de ser um cientista de dados líder](#)

[16.2.2 Desvantagens de ser um cientista de dados líder](#)

[16.2.3 Como se tornar um cientista de dados líder](#)

[16.3 Mudar para consultoria independente](#)

[16.3.1 Benefícios da consultoria independente](#)

[16.3.2 Desvantagens da consultoria independente](#)

[16.3.3 Como se tornar um cientista de dados líder](#)

[16.4 Como escolher seu caminho](#)

[16.5 Entrevista com Angela Bassa, chefe da ciência de dados, engenheira de dados e machine learning na iRobot](#)

[Resumo](#)

[Recursos dos capítulos 13–16](#)

[Livros](#)

[Blogs](#)

## **Epílogo**

## **Apêndice**

## **Perguntas da entrevista**

# Prefácio

“Como eu consigo um trabalho como o seu?”

Por sermos cientistas de dados experientes, constantemente nos fazem essa pergunta. Às vezes, diretamente, outras vezes de forma indireta, com perguntas sobre as decisões que tomamos em nossa carreira para chegar onde estamos. Por trás dessa pergunta, as pessoas parecem estar em uma luta constante, porque há tão poucos recursos para descobrir como tornar-se um cientista de dados ou evoluir na carreira. Muitos cientistas de dados procuram ajuda para suas carreiras e, muitas vezes, não encontram respostas claras.

Embora tenhamos escrito posts de blog com conselhos táticos sobre como lidar com momentos específicos em um trabalho de Ciência de Dados, tivemos dificuldade devido à falta de um texto definitivo que aborde amplamente o desenvolvimento de uma carreira em Ciência de Dados. Este livro foi escrito com o objetivo de auxiliar esses profissionais, ou seja, aqueles que têm conhecimento sobre Ciência de Dados e machine learning (aprendizado de máquina), mas não sabem por onde começar, e também aqueles que já estão no campo e querem saber como crescer.

A oportunidade de colaborar na elaboração deste trabalho foi prazerosa e sentimos que nossas formações e pontos de vista se complementaram e, assim, criamos um livro melhor para você. Nós somos:

- *Jacqueline Nolis* – bacharel e mestre em matemática e doutora em pesquisas operacionais. Quando comecei a trabalhar, o termo *ciência de dados* ainda não existia, e tive de descobrir o caminho enquanto a área estava se definindo. Atualmente, sou consultora e colaboro com empresas no desenvolvimento de equipes de ciência de dados.
- *Emily Robinson* – formada em ciência da decisão e mestre em administração. Depois de participar de um bootcamp de ciência de dados durante três meses em 2016, comecei a trabalhar em ciência de dados, especializando-me em testes A/B. Agora trabalho como cientista

de dados na Warby Parker, responsável por alguns dos maiores projetos da empresa.

Ao longo de nossas carreiras, criamos portfólios de projetos e passamos pelo estresse de nos adaptarmos a um novo emprego. Sentimos a angústia de sermos rejeitadas por empregos que queríamos e pelo contentamento de ver nossas análises impressionar positivamente as empresas. Enfrentamos problemas com difíceis parceiros de negócios e recebemos apoio de um mentor. Embora essas experiências tenham nos ensinado tanto em nossas carreiras, para nós o verdadeiro valor está em compartilhá-las com os demais.

Este livro foi pensado para ser um guia e responder às perguntas da carreira em ciência de dados, seguindo o trajeto que uma pessoa fará na carreira. Começamos desde o início da jornada: como obter competências básicas em ciências de dados e compreender como os empregos de fato são. Depois, tratamos sobre como conseguir um emprego e se adaptar. Abordamos como crescer na função e como fazer a transição até a gestão – ou para uma empresa nova. Nossa intenção é que este livro seja um recurso ao qual os cientistas de dados sigam, retornando à medida que atinjam novos marcos em suas carreiras.

Como o foco na carreira é muito importante para este livro, escolhemos não nos concentrarmos profundamente nos componentes técnicos da ciência de dados; não abordamos tópicos, como, por exemplo, escolher os hiperparâmetros de um modelo ou os pequenos detalhes dos pacotes Python. Na verdade, este livro não inclui uma única equação ou linha de código. Sabemos que muitos livros abordam esses tópicos; em vez disso, gostaríamos de discutir o conhecimento não técnico frequentemente negligenciado, mas igualmente importante, necessário para ter sucesso em ciência de dados.

Incluímos neste livro muitas experiências pessoais de renomados cientistas de dados. Ao final de cada capítulo, apresenta-se uma entrevista que descreve como um cientista de dados humano e real lidou pessoalmente com os conceitos que o capítulo aborda. Estamos extremamente satisfeitas com as respostas incríveis e detalhadas que todos os cientistas de dados entrevistados nos deram. Sentimos que os exemplos que dão de suas vidas



podem ensinar muito mais do que qualquer outra coisa que poderíamos escrever.

Outra decisão que tomamos ao escrever este livro foi a de trazer várias opiniões, ou seja, escolhemos intencionalmente nos concentrarmos nas lições que aprendemos como cientistas de dados profissionais em contato com a comunidade. Às vezes, fazemos declarações que nem todos podem concordar, como sugerir que você deva sempre escrever uma carta de apresentação ao se candidatar a uma vaga. Sentimos que o benefício de trazer pontos de vista nos quais acreditamos firmemente como sendo úteis aos cientistas de dados era mais importante do que tentar escrever algo que contivesse apenas verdades objetivas.

Esperamos que este livro seja um guia útil à medida que você avança na sua carreira em ciência de dados. Escrevemos este livro como se fosse o material que queríamos ter quando éramos aspirantes a cientistas de dados; esperamos que você fique satisfeito.

# Agradecimentos

Antes de mais nada, gostaríamos de agradecer aos nossos cônjuges, Michael Berkowitz e Heather Nolis. Sem eles, este livro não teria sido possível (e não apenas porque Michael escreveu um primeiro rascunho de algumas das seções, apesar de ser um profissional de bridge [jogo de cartas], e não um cientista de dados, ou porque Heather evangelizou metade do conteúdo sobre engenharia de machine learning).

Em seguida, queremos reconhecer a equipe da Manning, que nos orientou neste processo, melhorou o livro e o tornou possível. Um agradecimento especial à nossa editora, Karen Miller, que nos manteve na direção certa e coordenou todas as inúmeras peças em movimento.

Somos gratas a todos os revisores que leram o manuscrito em vários momentos e forneceram feedbacks detalhados e inestimáveis: Brynjar Smári Bjarnason, Christian Thoudahl, Daniel Berecz, Domenico Nappo, Geoff Barto, Gustavo Gomes, Hagai Luger, James Ritter, Jeff Neumann, Jonathan Twaddell, Krzysztof Jeźdrzejewski, Malgorzata Rodacka, Mario Giesel, Narayana Lalitanand Surampudi, Ping Zhao, Riccardo Marotti, Richard Tobias, Sebastian Palma Mardones, Steve Sussman, Tony M. Dubitsky e Yul Williams. Agradecemos também aos nossos amigos e familiares que leram o livro e apresentaram suas sugestões: Elin Farnell, Amanda Liston, Christian Roy, Jonathan Goodman e Eric Robinson. Suas contribuições ajudaram a formar este livro e deixaram-no o mais proveitoso possível aos nossos leitores.

Por fim, queremos agradecer a todos os nossos entrevistados dos capítulos: Robert Chang, Randy Au, Julia Silge, David Robinson, Jesse Mostipak, Kristen Kehrer, Ryan Williams, Brooke Watson Madubunwu, Jarvis Miller, Hilary Parker, Heather Nolis, Sade Snowden-Akintunde, Michelle Keim, Renee Teate, Amanda Casari e Angela Bassa. Além disso, agradecemos àqueles que contribuíram para as barras laterais do livro e para as perguntas de entrevista sugeridas para o Apêndice: Vicki Boykis,

Rodrigo Fuentealba Cartes, Gustavo Coelho, Emily Bartha, Trey Causey, Elin Farnell, Jeff Allen, Elizabeth Hunter, Sam Barrows, Reshama Shaikh, Gabriela de Queiroz, Rob Stamm, Alex Hayes, Ludamila Janda, Ayanthi G., Allan Butler, Heather Nolis, Jeroen Janssens, Emily Spahn, Tereza Iofciu, Bertil Hatt, Ryan Williams, Peter Baldrige e Hlynur Hallgrímsson. Todas essas pessoas deram perspectivas valiosas e, juntas, elas sabem muito mais do que poderíamos sonhar.

# Sobre este livro

*Construa uma Carreira em Ciência de Dados* foi escrito para ajudar o leitor a ingressar no campo da ciência de dados e a desenvolver uma carreira. O livro orienta sobre a função de um cientista de dados, como obter as competências de que precisa e as etapas para conseguir um emprego em ciência de dados. Depois de conseguir um emprego, este livro pode ajudá-lo a se aprimorar na função e, posteriormente, destacar-se comunidade de ciência de dados, bem como ser um experiente cientista de dados. Depois de ler este livro, você terá segurança para prosperar na sua carreira.

## Quem deve ler este livro

Este livro destina-se àqueles que ainda não ingressaram no campo da ciência de dados, mas que estão considerando o acesso, bem como a pessoas que estão nos primeiros anos da função. Os aspirantes a cientistas de dados aprenderão as competências de que necessitam para se tornarem cientistas de dados, enquanto os cientistas de dados em início de carreira aprenderão a se tornar mais experientes. Muitos dos tópicos do livro, como entrevistas e negociação de uma oferta, são recursos valiosos para retornar ao longo de qualquer carreira em ciência de dados.

## Como este livro está organizado: um roteiro

Este livro está dividido em quatro partes, organizadas em ordem cronológica de uma carreira em ciência de dados. A Parte I do livro, *Introdução à ciência de dados*, aborda o que é a ciência de dados e quais competências ela requer:

- O Capítulo 1 introduz a função de um cientista de dados e os diferentes tipos de empregos que têm esse nome.
- O Capítulo 2 apresenta cinco empresas de exemplo que têm cientistas de dados e mostra como a cultura e o tipo de cada empresa afeta os cargos

da ciência de dados.

- O Capítulo 3 mostra os diferentes caminhos que uma pessoa pode seguir para aprender as competências necessárias para ser um cientista de dados.
- O Capítulo 4 descreve como criar e compartilhar projetos para construir um portfólio em ciência de dados.

A Parte II do livro, *Como encontrar emprego em ciência de dados*, explica todo o processo de busca de emprego para cargos da ciência de dados:

- O Capítulo 5 percorre a busca por vagas abertas e como encontrar as que valem a pena investir.
- O Capítulo 6 explica como criar uma carta de apresentação e um currículo e, depois, adaptá-los para cada vaga.
- O Capítulo 7 fornece detalhes sobre o processo de entrevista e o que esperar dela.
- O Capítulo 8 trata sobre o que fazer depois que você recebe uma oferta, com foco em como negociá-la.

A Parte III do livro, *Adaptação em ciência de dados*, abrange os princípios básicos dos primeiros meses de um trabalho em ciência de dados:

- O Capítulo 9 mostra o que esperar dos primeiros meses de um trabalho em ciência de dados e como tirar o máximo proveito deles.
- O Capítulo 10 percorre o processo de realização de análises, que são componentes centrais da maioria das funções da ciência de dados.
- O Capítulo 11 concentra-se em colocar modelos de machine learning em produção, o que é necessário em cargos com mais base em engenharia.
- O Capítulo 12 explica como se comunicar com stakeholders – uma tarefa que os cientistas de dados têm de fazer mais do que a maioria das outras funções técnicas.

A Parte IV do livro, *Desenvolvimento da função em ciência de dados*, aborda tópicos para cientistas de dados mais experientes que estejam buscando progredir em suas carreiras:

- O Capítulo 13 descreve como lidar com falhas em projetos de ciência de dados.

- O Capítulo 14 mostra como fazer parte da comunidade de ciência de dados por meio de atividades, como falar e contribuir para o código aberto.
- O Capítulo 15 é um guia para a difícil tarefa de sair de um cargo em ciência de dados.
- O Capítulo 16 encerra o livro com as funções que os cientistas de dados podem obter à medida que avançam na estrada corporativa.

Finalmente, temos um Apêndice com mais de 30 perguntas de entrevista, respostas de exemplo e notas sobre o que a questão está tentando avaliar e o que faz dela uma boa resposta.

As pessoas que não são cientistas de dados devem começar no início do livro. Já aquelas que estão no campo podem começar em um capítulo posterior para orientá-las nos desafios que estão sendo enfrentados. Embora os capítulos sejam ordenados a fluir como uma carreira na ciência de dados, podem ser lidos fora de ordem de acordo com as necessidades.

Os capítulos encerram com entrevistas de cientistas de dados em vários setores que discutem como o tópico do capítulo tem aparecido na carreira. Os entrevistados foram selecionados em decorrência de suas contribuições para o campo da ciência de dados e os caminhos interessantes que tomaram ao se tornarem cientistas de dados.

# Sobre as autoras



## Emily Robinson

### Escrito por Jacqueline Nolis

Emily Robinson é uma cientista de dados brilhante na Warby Parker e trabalhou antes na DataCamp e na Etsy.

Conheci a Emily no Data Day Texas 2018, quando ela era uma das poucas pessoas que assistiram à minha palestra sobre ciência de dados no setor. No fim da minha palestra, ela levantou a mão e fez uma ótima pergunta. Para minha surpresa, uma hora mais tarde tínhamos trocado de lugar; eu a estava vendo calma e casualmente dando uma ótima palestra, enquanto eu estava ansiosamente esperando para levantar minha mão e fazer uma pergunta. Naquele dia, sabia que ela era uma cientista de dados inteligente e esforçada. Alguns meses mais tarde, quando chegou a hora de encontrar alguém para a coautoria do livro, ela estava no topo da minha lista. Quando enviei um email perguntando se estaria interessada, pensei que havia chances de ela dizer não; provavelmente ela estaria fora do meu alcance.

Trabalhar com Emily neste livro foi ótimo. Ela é excessivamente atenta às dificuldades dos cientistas de dados em início de carreira e tem a

capacidade de compreender com clareza o que é importante. Está sempre fazendo seu trabalho e, de alguma maneira, também consegue encaixar postagens extras no blog. Agora, já a tendo visto em mais conferências e eventos sociais, vi como se comunica com muitos cientistas de dados e faz com que todos se sintam bem-vindos. É também especialista em testes e experimentos A/B, e que por acaso é a área em que ela está trabalhando no momento; poderia escolher qualquer outra parte da ciência de dados e ser uma especialista na área, se quisesse.

Minha única decepção é que estou escrevendo essas palavras sobre ela no fim da elaboração do livro. Encerrado esse livro alguém além de mim terá a próxima oportunidade de colaborar com ela.

## **Jacqueline Nolis**

### **Escrito por Emily Robinson**

Sempre que alguém me pergunta se eu recomendaria escrever um livro, digo: “Só se você o fizer em coautoria”. Mas não é só isso. É mais: “Apenas se o fizer com uma coautora tão divertida, calorosa, generosa, inteligente, experiente e atenciosa como a Jacqueline”. Não tenho certeza de como é trabalhar com um coautor “normal”, porque a Jacqueline sempre foi fantástica, e sinto-me incrivelmente sortuda por ter conseguido dividir com ela este projeto.

Seria fácil alguém tão talentosa como a Jacqueline ser intimidativa. Ela tem doutorado em engenharia industrial, ganhou 100 mil dólares por vencer a terceira temporada do reality show *King of the Nerds*, foi diretora de análises e começou sua própria empresa de consultoria bem-sucedida. Proferiu palestras em conferências por todo o país e sua alma mater regularmente solicita que ela oriente a respeito de carreira alunos da graduação em matemática (a formação dela). Quando palestrou em uma conferência online, os elogios sobre sua apresentação inundaram o bate-papo, como “a melhor até agora”, “apresentação excelente”, “realmente útil” e “apresentação excelente e dinâmica”. A Jacqueline nunca deixa alguém se sentir inferior ou inoportuno por não saber algo; em vez disso, ela ama tornar acessíveis conceitos difíceis, como em sua grande



apresentação chamada de “deep learning não é difícil, eu prometo”.

Sua vida pessoal é igualmente impressionante: ela tem uma casa maravilhosa e vibrante em Seattle com esposa, filho, dois cães e três gatos. Espero que ela também possa um dia adotar uma certa coautora para preencher os poucos espaços vazios. Ela e sua esposa, Heather, inclusive apresentaram a uma plateia com milhares de pessoas ansiosas para ouvir sobre como usaram a linguagem R na implantação de modelos de machine learning para produção na T-Mobile. Elas provavelmente também têm a melhor história do mundo sobre como se conheceram: elas se encontraram no show mencionado antes, *King of the Nerds*, no qual Heather também concorria.

Sou muito grata à Jacqueline, que poderia ganhar muito mais dinheiro por muito menos, fazendo qualquer outra coisa além de escrever este livro comigo. Espero que nosso trabalho incentive os cientistas de dados aspirantes e em início de carreira a contribuírem para nossa comunidade, como ela fez.

# Sobre a ilustração da capa

## Saint-Sauver

A imagem na capa de *Construa uma Carreira em Ciência de Dados* chama-se “Femme de l’Aragon” ou “Mulher de Aragão”. A ilustração foi retirada de uma coleção de trajes de vários países por Jacques Grasset de Saint-Sauveur (1757-1810), intitulada *Costumes de Différents Pays*, publicada na França em 1797. Cada ilustração é desenhada com precisão e colorida à mão. A rica variedade da coleção de Grasset de Saint-Sauveur lembra-nos claramente como as cidades e regiões do mundo eram culturalmente diferentes há apenas 200 anos. Isoladas umas das outras, as pessoas falavam dialetos e línguas diferentes. Nas ruas ou no campo, era fácil identificar onde viviam e qual era seu comércio ou posição social apenas pelas vestimentas.

A forma como nos vestimos mudou desde então, e a diversidade de cada região, tão rica na época, perdeu-se. Agora é difícil distinguir os habitantes de diferentes continentes, muito menos de cidades, regiões ou países distintos. Talvez tenhamos trocado a diversidade cultural por uma vida pessoal mais variada – certamente por uma vida tecnológica mais variada e acelerada.

Em tempos nos quais é difícil distinguir um livro de informática de outro, a Manning celebra a inventividade e a iniciativa do setor de informática com capas de livros baseadas na rica diversidade da vida regional de dois séculos atrás, trazida de volta à vida pelas imagens de Grasset de Saint-Sauveur.

# PARTE I

## Introdução à ciência de dados

Se você fizer uma pesquisa no Google sobre *como se tornar um cientista de dados*, provavelmente irá se deparar com uma lista de competências, desde modelagem estatística até programação em Python, além de comunicação eficaz e fazer apresentações. Uma vaga de emprego pode descrever uma função que está próxima à de um estatístico, mas outro empregador está à procura de alguém que tenha um mestrado em ciência da computação. Ao buscar maneiras de aprender essas competências, você encontrará opções que vão desde cursar um mestrado, realizar um bootcamp e até começar a fazer análise de dados no seu trabalho de agora. Em conjunto, todas essas combinações de caminhos podem parecer intransponíveis, especialmente para pessoas que ainda não têm certeza se querem ser cientistas de dados.

A boa notícia é que não há um único cientista de dados sequer que tenha todas essas competências. Eles partilham uma base de conhecimento, mas cada um tem suas próprias especialidades, a ponto de que muitos não poderiam trocar os cargos entre si. A primeira parte deste livro foi concebida para auxiliar o leitor a compreender quais são esses tipos de cientistas de dados e como tomar as melhores decisões para iniciar sua carreira. Ao final, você terá as competências e a compreensão para iniciar sua busca por vagas de emprego.

O Capítulo 1 abrange os princípios básicos da ciência de dados, incluindo as competências necessárias para o trabalho e os diferentes tipos de cientistas de dados. O Capítulo 2 entra em detalhes sobre o papel de um cientista de dados em cinco tipos de empresas com o objetivo de ajudá-lo a entender melhor como será o trabalho. O Capítulo 3 trata dos caminhos para aprender essas competências necessárias para ser um cientista de dados, assim como as vantagens e desvantagens de cada um. Por fim, o Capítulo 4 trata sobre como criar um portfólio de projetos de ciência de dados para

obter experiência prática fazendo ciência de dados e mostrá-lo a possíveis empregadores.

# CAPÍTULO 1

## O que é a ciência de dados?

Este capítulo abrange:

- As três principais áreas da ciência de dados
- Os diferentes tipos de vagas em ciência de dados

“O trabalho mais sexy do século XXI”. “O melhor trabalho nos EUA”. Cientista de dados, um título que sequer existia antes de 2008, agora é o cargo para o qual os empregadores não param de contratar e que os candidatos se esforçam para ocupar. Há uma boa razão para o entusiasmo: a ciência de dados é um campo em crescimento enorme, com um salário base na média anual de mais de US\$ 100 mil nos Estados Unidos em 2019 (<http://mng.bz/XpMp>). Em uma boa empresa, os cientistas de dados desfrutam de muita autonomia e estão constantemente aprendendo coisas novas. Utilizam suas competências para resolver problemas significativos, como trabalhar com médicos para analisar ensaios clínicos de medicamentos, ajudar uma equipe esportiva a escolher seus novos integrantes ou redesenhar o modelo de preços para um negócio de widgets. Por fim, como discutiremos no Capítulo 3, não há uma única maneira de se tornar um cientista de dados. As pessoas vêm de todas as áreas, por isso não se limita ao que se estudou na universidade.

Mas nem todos os cargos na ciência de dados são perfeitos. Tanto as empresas como os candidatos podem ter expectativas não realistas. As empresas novas em ciência de dados podem imaginar que alguém pode resolver todos os problemas com dados, por exemplo. Quando um cientista de dados é finalmente contratado, ele pode ser confrontado com uma lista interminável de pedidos. Eles podem ter a tarefa de implementar imediatamente um sistema de machine learning (aprendizado de máquina) quando nenhum trabalho tiver sido feito para preparar ou limpar os dados.

Pode ser que não haja ninguém para guiá-los ou até mesmo entender os problemas que enfrentam. Discutiremos esses problemas em maior profundidade nos capítulos 5 e 7, onde auxiliaremos você a evitar ingressar em empresas que provavelmente não sejam uma boa opção para um novo cientista de dados. No Capítulo 9, aconselharemos o que fazer se acabar em uma situação difícil.

Por outro lado, os candidatos podem pensar que nunca haverá um momento tranquilo na nova carreira. Podem esperar que os stakeholders sigam suas recomendações rotineiramente, que os engenheiros de dados possam corrigir de imediato quaisquer problemas de qualidade de dados e que eles tenham os recursos de computação mais rápidos disponíveis para implementar os modelos. Na realidade, os cientistas de dados investem muito tempo limpando e preparando dados, assim como gerenciando as expectativas e prioridades de outras equipes. Os projetos nem sempre saíram conforme o esperado. A alta gerência pode fazer promessas não realistas aos clientes sobre o que seus modelos de ciência de dados podem oferecer. A função principal de uma pessoa pode ser trabalhar com um sistema de dados arcaico, que é impossível automatizar e requer horas de trabalho exaustivo a cada semana apenas para limpar os dados. Os cientistas de dados podem notar muitos erros estatísticos ou técnicos em análises herdadas, com consequências reais, sem que ninguém se preocupe, mas eles se encontram tão sobrecarregados com trabalho que não têm tempo para tentar resolvê-los. Ou pode ser solicitado a um cientista de dados que prepare relatórios que sustentem o que a alta gerência já decidiu, fazendo com que se preocupem se serão demitidos caso deem uma resposta diferente.

Este livro está aqui para guiá-lo no processo de se tornar um cientista de dados e desenvolver sua carreira. Queremos garantir que você conheça a parte boa de ser um cientista de dados e evite a maioria das armadilhas. Talvez esteja trabalhando em um campo adjacente, como análise de marketing, e se perguntando como fazer para mudar de área. Ou talvez já seja um cientista de dados, mas está buscando um novo trabalho porque acha que não abordou bem a sua primeira procura por emprego na área. Ou, ainda, quer continuar sua carreira palestrando em conferências,

contribuindo para o código aberto ou se tornando um consultor independente. Seja qual for seu nível, estamos confiantes de que este livro lhe será útil.

Nos primeiros quatro capítulos, abordamos as principais oportunidades para aprender competências em ciências de dados e construir um portfólio visando contornar o paradoxo de necessitar de experiência para conseguir experiência. A Parte II mostra como escrever uma carta de apresentação e elaborar o currículo que resultará em uma entrevista, além de como construir uma rede de contatos para conseguir uma recomendação. Tratamos de estratégias de negociação que pesquisas demonstraram que farão você conseguir a melhor oferta possível.

Quando estiver em um cargo de ciência de dados, você redigirá análises, trabalhará com stakeholders e talvez até mesmo colocará um modelo em produção. A Parte III ajuda a compreender como todos esses processos se parecem e como se preparar para uma trajetória de sucesso. Na Parte IV, há estratégias para se recuperar quando um projeto inevitavelmente falhar. E quando estiver pronto, estamos aqui para guiá-lo na decisão de onde levar sua carreira: ir para a gerência, continuar a ser colaborador individual ou até mesmo consultor independente.

Antes de começar essa jornada, porém, você precisa entender claramente o que são cientistas de dados e o que eles fazem. A ciência de dados é um campo amplo que abrange muitos tipos de trabalho e, quanto melhor entender as diferenças entre essas áreas, mais você pode se desenvolver nelas.

## **1.1 O que é a ciência de dados?**

A *ciência de dados* é a prática de usar dados para tentar entender e resolver problemas do mundo real. Esse conceito não é exatamente novo; as pessoas têm analisado números e tendências de vendas desde a invenção do zero. Na última década, entretanto, ganhamos acesso a exponencialmente mais dados do que já existiam. O advento dos computadores tem ajudado na geração de todos esses dados, mas a computação também é nossa única maneira de processar esse grande volume de informações. Com códigos,

um cientista de dados pode transformar ou agregar dados, executar análises estatísticas ou treinar modelos de machine learning. O resultado desse código pode ser um relatório ou um dashboard para consumo humano ou pode ser um modelo de machine learning que será implementado para ser executado de forma contínua.

Se uma empresa de varejo estiver tendo problemas para decidir onde colocar uma nova loja, por exemplo, ela pode chamar um cientista de dados para fazer uma análise. O cientista de dados poderia olhar os dados históricos dos locais para onde os pedidos online são enviados para entender onde está a demanda do cliente. Também pode combinar esses dados de localização de clientes com informações demográficas e de renda para essas localidades a partir de registros do censo. Com esses conjuntos de dados, ele poderiam encontrar o local ideal para a nova loja e criar uma apresentação com o Microsoft PowerPoint, encaminhando ao vice-presidente de operações de varejo da empresa.

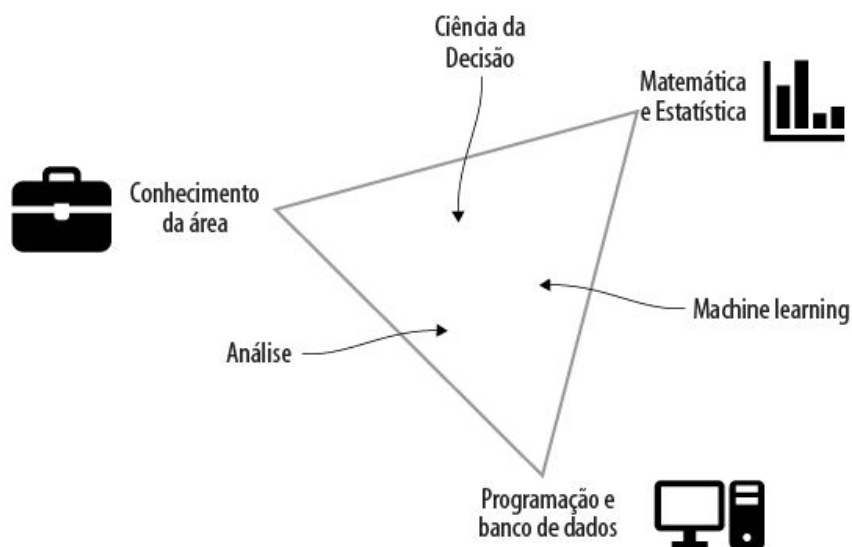
Em outra situação, essa mesma empresa de varejo pode querer aumentar os pedidos online, recomendando artigos aos clientes durante a compra. Um cientista de dados poderia carregar dados históricos de pedidos online e criar um modelo de machine learning que, com base nos artigos do carrinho de compras, prediz o melhor artigo a ser sugerido ao comprador. Depois de criar esse modelo, o cientista de dados trabalharia com a equipe de engenharia da empresa para que, cada vez que um cliente faz compras, o novo modelo de machine learning entregue os itens recomendados.

Quando muitas pessoas começam a olhar para a ciência de dados, um desafio que enfrentam está na quantidade surpreendente de coisas que precisam aprender, como programação (mas qual linguagem?), estatística (mas quais métodos são mais importantes na prática e quais são, em grande parte, acadêmicos?), machine learning (mas quão diferente é o machine learning da estatística ou da IA?) e o conhecimento sobre o setor em que querem trabalhar (mas e se não souberem onde querem trabalhar?). Além disso, eles precisam aprender competências corporativas, como comunicar resultados de forma eficaz a públicos que vão desde outros cientistas de dados ao CEO. Essa ansiedade pode ser exacerbada por vagas de emprego que pedem doutorado, vários anos de experiência em ciência de dados e



experiência em diversos métodos estatísticos e de programação. Como aprender todas essas competências? Com quais deve começar? Quais são os princípios básicos?

Se tiver investigado áreas diferentes da ciência de dados, é possível que esteja familiarizado com o famoso diagrama de Venn da ciência de dados de Drew Conway. Na opinião de Conway (no momento da criação do diagrama), as ciências de dados caíram na intersecção dos conhecimentos matemáticos e estatísticos, dos conhecimentos especializados em um domínio e das competências de hacking (isto é, programação). Essa imagem é muito utilizada como o alicerce da definição do que é um cientista de dados. Do nosso ponto de vista, os componentes da ciência de dados são ligeiramente diferentes do que ele propôs (Figura 1.1).



*Figura 1.1 – As competências combinadas para fazer ciência de dados e como se combinam para diferentes funções.*

Mudamos o diagrama de Venn original de Conway para um triângulo porque não se trata de ter uma competência ou não, mas, sim, de você possuí-la de forma diferente das outras pessoas no campo. Embora seja verdade que todas as três competências sejam fundamentais e que é necessário ter cada uma delas em certo nível, não é necessário ser um especialista em todas elas. Colocamos dentro do triângulo diferentes tipos de especialidades de ciência de dados. Essas especialidades nem sempre correspondem aos nomes dos cargos, e, mesmo quando é o caso, as

empresas às vezes querem dizer coisas diferentes.

O que significa cada um desses componentes?

### 1.1.1 Matemática/estatística

Em um nível básico, o conhecimento matemático e estatístico trata da alfabetização ou literacia de dados. Dividimos essa alfabetização em três níveis de conhecimento:

- *Essas técnicas existem* – se não souber que algo é possível, não poderá usá-lo. Se um cientista de dados estivesse tentando agrupar clientes semelhantes, sabendo que os métodos estatísticos (chamados de *agrupamento ou clustering*) podem fazer isso, essa seria a primeira etapa.
- *Como aplicar as técnicas* – embora um cientista de dados possa conhecer muitas técnicas, ele também precisa ser capaz de entender as complexidades de aplicá-las – não apenas como escrever código para aplicar os métodos, mas também como configurá-los. Se o cientista de dados quiser usar um método como o clustering  $k$ -means para agrupar os clientes, ele precisará entender como fazer um clustering  $k$ -means em uma linguagem de programação como R ou Python. Ele também precisaria entender como ajustar os parâmetros do método, por exemplo, escolhendo quantos grupos criar.
- *Como escolher as técnicas que devem ser utilizadas* – uma vez que tantas técnicas possíveis podem ser utilizadas na ciência de dados, é importante que o cientista de dados consiga avaliar rapidamente se uma técnica funcionaria bem. Em nosso exemplo de agrupamento de clientes, mesmo depois que o cientista de dados se concentra em clustering, ele tem que considerar dezenas de métodos e algoritmos diferentes. Em vez de tentar cada método, precisa ser capaz de excluir métodos com rapidez e se concentrar em apenas alguns.

Esses tipos de competências são usados constantemente em uma função de ciência de dados. Para considerar um exemplo diferente, suponha que você trabalha em uma empresa de ecommerce. Seu parceiro de negócios pode estar interessado nos países que têm o valor médio mais alto de pedidos. Se

tiver os dados disponíveis, essa pergunta é fácil de responder. Mas em vez de simplesmente apresentar essas informações e deixar seu cliente tirar as próprias conclusões, é possível se aprofundar. Se tiver um pedido do país A por US\$ 100 e mil pedidos do país B que tenham uma média de US\$ 75, é correto que o país A tem o valor médio de pedido mais elevado. Porém, você estaria confiante em dizer que significa que seu parceiro de negócios deve definitivamente investir em publicidade no país A para aumentar o número de pedidos? Provavelmente não. Você tem apenas um ponto de dados para o país A e talvez seja uma exceção. Se o país A tivesse 500 pedidos, seria possível usar um teste estatístico para ver se o valor do pedido era significativamente diferente, o que quer dizer que se realmente não houvesse diferença entre A e B nessa medida seria improvável ver a diferença. Nesse exemplo, foram feitas muitas avaliações diferentes sobre quais abordagens eram sensatas, o que deveria ser considerado e quais resultados foram considerados pouco importantes.

### **1.1.2 Bancos de dados/programação**

*A programação e os bancos de dados* referem-se à capacidade de extrair dados dos bancos de dados da empresa e de escrever códigos claros, eficientes e sustentáveis. Essas competências são de muitas maneiras similares às que um desenvolvedor de software precisa saber, exceto que os cientistas de dados têm que escrever um código que faz uma análise aberta em vez de produzir um resultado predefinido. Os dados de cada empresa são únicos, portanto, não é necessário um conjunto de competências técnicas específico para um cientista de dados. No entanto, em geral, é preciso saber como obter dados de um banco de dados e como limpar, manipular, resumir, visualizar e compartilhar dados.

Na maioria dos cargos de ciência de dados, R ou Python é a linguagem principal. R é uma linguagem de programação que tem suas raízes na estatística, portanto, é geralmente mais forte para análise e modelagem estatísticas, além da visualização e da geração de relatórios com resultados. Python é uma linguagem de programação que começou como uma linguagem geral de desenvolvimento de software e se tornou extremamente conhecida na ciência de dados. Python é conhecida por ser melhor do que R

ao trabalhar com grandes conjuntos de dados, fazer machine learning e acionar algoritmos em tempo real (como os mecanismos de recomendação da Amazon). Mas graças ao trabalho de muitos colaboradores, as capacidades das duas linguagens estão agora quase em paridade. Os cientistas de dados estão usando R com sucesso para fazer modelos de machine learning, que são rodados milhões de vezes por semana e estão fazendo análises estatísticas claras e apresentáveis em Python.

R e Python são as linguagens mais conhecidas para a ciência de dados por algumas razões:

- Elas são livres e de código aberto, o que significa que muitas pessoas, não apenas uma empresa ou um grupo, contribuem com o código que você pode usar. Elas têm muitos pacotes ou *bibliotecas* (conjuntos de código) para fazer a coleta, manipulação, visualização, análise estatística e machine learning de dados.
- Como cada linguagem tem um grande número de usuários, é fácil para os cientistas de dados encontrar ajuda quando se defrontam com problemas. Embora algumas empresas ainda usem SAS, SPSS, STATA, MATLAB ou outros programas pagos, muitas delas estão começando a mudar para R ou Python.

Ainda que a maioria das análises de ciência de dados seja feita em R ou Python, em geral será preciso trabalhar com um banco de dados para obter os dados. É aqui que a linguagem SQL entra. SQL é a linguagem de programação que a maioria dos bancos de dados usa para manipular dados dentro deles ou para extraí-los. Considere um cientista de dados que queira analisar as centenas de milhões de registros de pedidos de clientes em uma empresa para prever como os pedidos por dia serão alterados ao longo do tempo. Primeiro, eles provavelmente escreveriam uma consulta SQL para obter o número de pedidos de cada dia. Então levariam essas contagens diárias de pedidos e rodariam uma previsão estatística em R ou Python. Por essa razão, SQL é extremamente usada na comunidade de ciência de dados, sendo difícil chegar muito longe sem ter conhecimento dela.

Outra especialidade básica é usar o *controle de versão* – um método de controlar como o código muda ao longo do tempo. O controle de versão permite que você armazene os arquivos, reverta-os para um período anterior

e veja quem alterou o arquivo, como e quando. Essa competência é extremamente importante para a ciência de dados e a engenharia de software, porque se alguém mudar acidentalmente um arquivo que quebra seu código, é bom ter a capacidade de revertê-lo ou ver o que mudou.

Git é de longe o sistema mais comumente usado para controle de versão e é com frequência empregado em conjunto com GitHub, um serviço de hospedagem baseado na Web para Git. Git permite que você salve (*commit*) suas mudanças, e também veja todo o histórico do projeto e como ele mudou em cada commit. Se duas pessoas estiverem trabalhando no mesmo arquivo separadamente, o Git garante que nenhum trabalho seja excluído ou substituído acidentalmente. Em muitas empresas, sobretudo naquelas com boas equipes de engenharia, será preciso usar o Git se quiser compartilhar o código ou colocar algo em produção.

**Você pode ser um cientista de dados sem programar?**

É possível fazer muitos trabalhos de dados usando somente o Excel, o Tableau ou outras ferramentas de inteligência de negócios que tenham interfaces gráficas. Embora você não esteja escrevendo código, essas ferramentas afirmam ter muitas das mesmas funcionalidades que as linguagens como R ou Python, às vezes usadas por muitos cientistas de dados. Mas elas podem ser um kit completo de ferramentas para a ciência de dados? Dizemos que não. Na prática, pouquíssimas empresas têm uma equipe de ciência de dados em que não seria preciso programar. Mas mesmo que não fosse o caso, a programação tem vantagens em relação à utilização dessas ferramentas.

A primeira vantagem da programação é a capacidade de se reproduzir os dados. Quando você escreve código em vez de usar software de apontar e clicar, poderá rodá-lo novamente sempre que seus dados forem alterados, sejam eles alterados todos os dias ou a cada seis meses. Essa vantagem também se vincula ao controle de versão: em vez de renomear seu arquivo sempre que o código for alterado, é possível manter um arquivo, mas ver todo o histórico.

A segunda vantagem é a flexibilidade. Se o Tableau não tiver um tipo de gráfico disponível, por exemplo, você não poderá criá-lo. Entretanto, com a programação, é possível escrever seu próprio código para fazer algo que os criadores e mantenedores de uma ferramenta nunca pensaram.

A terceira e última vantagem das linguagens de código aberto, como Python e R, é a contribuição da comunidade. Milhares de pessoas criam *pacotes* e os publicam abertamente em GitHub e/ou CRAN (para R) e pip (para Python). Você pode fazer o download desse código e utilizá-lo para seus próprios problemas, independentemente de uma empresa ou grupo de pessoas para adicionar recursos.

### 1.1.3 Entendimento dos negócios

*Qualquer tecnologia suficientemente avançada é indistinguível da magia.*

Arthur C. Clarke

As empresas têm, para dizer de forma educada, uma compreensão variável da forma como a ciência de dados funciona. Muitas vezes, a gestão apenas quer que algo seja feito e recorre aos unicórnios da ciência de dados para fazer com que algo aconteça. Uma competência central na ciência de dados é saber como traduzir uma situação corporativa em uma questão de dados, encontrar a resposta dos dados e finalmente entregar a resposta do negócio. Um empresário pode perguntar, por exemplo: “Por que nossos clientes estão nos deixando?”. Mas não há um pacote do Python “por-que-os-clientes-estão-nos-deixando” que você pode importar – cabe-lhe deduzir a maneira de responder a essa pergunta com dados.

O entendimento dos negócios é onde seus ideais de ciência de dados atendem às práticas do mundo real. Não é suficiente querer uma informação específica sem saber como os dados são armazenados e atualizados em sua empresa. Se sua empresa for um serviço de subscrição, onde estão os dados? Se alguém alterar a subscrição, o que acontece? A linha desse assinante é atualizada ou outra linha é adicionada à tabela? Você precisa contornar quaisquer erros ou inconsistências nos dados? Se não souber as respostas a essas perguntas, não poderá dar uma resposta precisa a uma pergunta básica como “quantos assinantes tínhamos em 2 de março de 2019?”.

O entendimento do negócio ajuda você a saber também quais perguntas fazer. Quando stakeholders perguntam “o que devemos fazer a seguir?”, é como se perguntassem “por que não temos mais dinheiro?”. Esse tipo de

pergunta gera mais perguntas. Desenvolver um entendimento do negócio principal (bem como das personalidades envolvidas) pode ajudá-lo a analisar melhor a situação. Você poderia dar continuidade perguntando: “Para qual linha de produtos estão buscando orientação?” ou “Gostariam de ver mais participação de um determinado setor do nosso público?”.

Outra parte do entendimento do negócio é desenvolver competências gerais do negócio, como poder adaptar as apresentações e os relatórios a públicos diferentes. Às vezes, você discutirá uma metodologia melhor com uma sala cheia de doutores em estatística, e outras vezes estará na frente do vice-presidente que não tem aula de matemática há mais de 20 anos. Você precisa informar seu público sem nivelar por baixo ou complicar demais.

Por fim, à medida que você se torna mais experiente, parte do seu trabalho é identificar onde a empresa pode se beneficiar da ciência de dados. Se você quiser criar um sistema de previsão para sua empresa, mas nunca teve suporte da gestão, talvez fazer parte da equipe de gestão ajude a resolver o problema. Um cientista de dados sênior estará à procura de locais para implementar machine learning, já que conhece suas limitações e capacidades, bem como os tipos de tarefas que se beneficiariam da automação.

## **A ciência de dados desaparecerá?**



Por trás dessa pergunta sobre se a ciência de dados estará por aqui em uma década ou mais encontramos duas preocupações básicas: que o trabalho se tornará automatizado e que a ciência de dados está na moda, e esta bolha do mercado de trabalho irá estourar em algum momento.

É verdade que algumas partes da ciência de dados podem ser automatizadas. O AutoML (Automated Machine Learning) pode comparar o desempenho de diferentes modelos e executar algumas partes da preparação de dados (como variáveis de escala). Mas essas tarefas são apenas uma pequena parte do processo da ciência de dados. Muitas vezes, você precisará criar os dados sozinho, por exemplo; é muito raro ter dados perfeitamente limpos esperando por você. Além disso, a criação dos dados geralmente envolve falar com outras pessoas, como pesquisadores ou engenheiros de experiência do usuário, os quais conduzirão a pesquisa ou registrarão as ações do usuário que podem conduzir a análise.

Quanto à possibilidade de estourar a bolha do mercado de trabalho, uma boa comparação é a engenharia de software na década de 1980. À medida que os computadores se tornaram mais baratos, mais rápidos e mais comuns, havia preocupações de que logo um computador poderia fazer tudo e que não haveria necessidade de programadores. Mas o oposto aconteceu, e agora há mais de 1,2 milhões de engenheiros de software nos Estados Unidos (<http://mng.bz/MOPo>). Embora títulos como webmaster tenham desaparecido, mais pessoas estão trabalhando no desenvolvimento, na manutenção e na melhoria de sites.

Acreditamos que haverá mais especialização na ciência de dados, o que pode levar ao desaparecimento do título geral de cientista de dados, mas muitas empresas ainda estão nas fases iniciais do aprendizado sobre como aprimorar a ciência de dados, havendo muito trabalho a ser feito.

## **1.2 Tipos diferentes de vagas na ciência de dados**

É possível misturar e combinar as três competências centrais da ciência de dados (tratadas na Seção 1.1) em diversos trabalhos, todas elas com alguma justificativa para ter cientista de dados como título. Do nosso ponto de vista, essas competências misturam-se de três formas principais: análise, machine learning e ciência da decisão (decision science). Cada uma dessas áreas serve a um propósito diferente para a empresa e fundamentalmente serve para algo diferente.

Ao procurar emprego na área da ciência de dados, preste menos atenção aos nomes dos cargos e muito mais às descrições de trabalho e ao que perguntam nas entrevistas. Veja as áreas das pessoas em cargos na ciência de dados, seus empregos anteriores e quais são suas formações. Você verá que as pessoas em funções similares têm cargos totalmente diferentes ou que as pessoas que têm o mesmo cargo de cientista de dados fazem coisas totalmente diferentes. Como falamos neste livro sobre diferentes tipos de empregos em ciência de dados, lembre-se de que os cargos reais usados nas empresas podem variar.

### **1.2.1 Análise**

Um *analista* pega dados e os coloca diante das pessoas certas. Depois que uma empresa ajusta os objetivos anuais, você pode colocar aqueles objetivos em um dashboard de modo que a gestão possa acompanhar o progresso a cada semana. Também é possível criar recursos que permitam que os gerentes desconstruam facilmente os números por país ou tipo de produto. Esse trabalho envolve muita limpeza e preparação de dados, mas em geral menos trabalho para interpretar os dados. Embora seja possível identificar e corrigir problemas de qualidade de dados, a pessoa mais importante a tomar decisões com esses dados é o parceiro de negócios.

Assim, o trabalho de um analista é pegar os dados da empresa, formatá-los e organizá-los de forma eficaz e entregá-los a outros.

Como o papel do analista não envolve muita estatística nem machine learning, algumas pessoas e empresas considerariam esse papel fora do campo da ciência de dados. Mas grande parte do trabalho, como conceber visualizações significativas e decidir transformações de dados específicas, requer as mesmas competências utilizadas nos outros tipos de funções da ciência de dados. Um analista pode receber como tarefa “criar um dashboard automatizado que mostre como nosso número de assinantes está mudando ao longo do tempo e que nos permita filtrar os dados apenas para ajustar assinantes de produtos específicos ou em regiões geográficas específicas”. O analista teria que encontrar os dados apropriados na empresa, descobrir como transformá-los de forma apropriada (por exemplo, mudando-os de novas assinaturas diárias para semanais) e, então, criar um conjunto significativo de dashboards visualmente atraente e automaticamente atualizado a cada dia sem erros.

Regra curta: um analista cria *dashboards e relatórios que fornecem dados*.

### **1.2.2 Machine learning**

Um *engenheiro de machine learning* desenvolve modelos de machine learning e coloca-os em produção, onde funcionam continuamente. Ele pode otimizar o algoritmo de classificação para os resultados de pesquisa de um site de ecommerce, criar um sistema de recomendação ou monitorar um modelo na produção para garantir que o desempenho não tenha ficado ruim desde a implantação. Um engenheiro de machine learning investe menos tempo em coisas como criar visualizações que convencerão as pessoas de algo e mais tempo fazendo o trabalho de programação da ciência de dados.

Uma grande diferença entre essa função e outros tipos de cargos na ciência de dados é que o resultado do trabalho está principalmente voltado para máquinas. Você pode criar modelos de machine learning que se transformem em interfaces de programação de aplicativos (APIs) para outras máquinas, por exemplo. Em muitos aspectos, você estará mais perto de um desenvolvedor de software do que de outras funções da ciência de dados. Embora seja bom que todo cientista de dados siga as melhores

práticas de programação, como engenheiro de machine learning, você precisa fazer isso. Seu código deve ser eficaz, testado e escrito para que outras pessoas possam trabalhar com ele. Por essa razão, muitos engenheiros de machine learning são da ciência da computação.

Na função de engenheiro de machine learning, pode ser necessário criar um modelo de machine learning que possa, em tempo real, prever a probabilidade de um cliente do site terminar seu pedido. O engenheiro de machine learning teria que encontrar dados históricos na empresa, treinar um modelo de machine learning nele, transformar esse modelo em uma API e, então, implantar a API para que o site possa executar o modelo. Se esse modelo para de funcionar por algum motivo, o engenheiro de machine learning será chamado para corrigi-lo.

Regra curta: um engenheiro de machine learning cria *modelos que são executados continuamente*.

### **1.2.3 Ciência de decisão**

Um *cientista de decisão* transforma os dados brutos de uma empresa em informações que a auxiliam na tomada de decisões. Esse trabalho depende de ter uma compreensão profunda dos diferentes métodos matemáticos e estatísticos e da familiaridade com a tomada de decisões empresariais. Além disso, os cientistas da decisão têm que conseguir fazer visualizações e tabelas atraentes, de modo que as pessoas não técnicas com quem compartilham as informações compreendam a análise. Embora um cientista de decisão programe muito, o trabalho dele em geral é executado apenas uma vez para realizar uma análise específica, a fim de que possam evitar ter um código ineficiente ou difícil de manter.

Um cientista de decisão deve entender as necessidades das outras pessoas dentro da empresa e descobrir como gerar informações construtivas. Um diretor de marketing, por exemplo, pode pedir a um cientista de decisão que o ajude a decidir quais tipos de produtos devem ser destacados no guia de presentes de férias da empresa. O cientista de decisão pode investigar quais produtos venderam bem sem serem incluídos no guia de presentes, falar com a equipe de pesquisa do usuário sobre a realização de uma pesquisa e usar princípios de ciência comportamental para fazer uma análise para

encontrar os itens ideais a serem sugeridos. O resultado provavelmente será uma apresentação ou relatório em PowerPoint a ser compartilhado com gerentes de produto, vice-presidentes e outros empresários.

Um cientista de decisão usa com frequência seu conhecimento de estatística para auxiliar a empresa na tomada de decisões em momentos de incerteza. Um cientista de decisão poderia ser responsável por gerenciar o sistema de análise de experimentos da empresa, por exemplo. Muitas empresas fazem experimentos online, ou testes A/B, para medir se uma mudança é eficaz. Essa mudança poderia ser tão simples como adicionar um botão novo ou tão complicada como mudar o sistema de classificação dos resultados de pesquisa ou redesenhar completamente uma página. Durante um teste A/B, os visitantes são aleatoriamente designados para uma de duas ou mais condições, por exemplo, metade para a versão antiga da página inicial, que é o *controle*, e metade para a nova versão, que é o *tratamento*. Em seguida, as ações dos visitantes após a entrada no experimento são comparadas para ver se os participantes no tratamento têm uma taxa mais elevada de ações desejáveis, como a compra de produtos.

Devido à aleatoriedade, é raro que as métricas no controle e tratamento sejam exatamente as mesmas. Suponha que você jogou cara ou coroa com duas moedas, e uma deu cara 52 vezes em 100 e outra 49 vezes de 100. Você concluiria que a primeira moeda é mais provável de dar cara? Claro que não! Mas um parceiro de negócios pode olhar para um experimento, ver que a taxa de conversão é de 5,4% no controle e 5,6% no tratamento e declarar o tratamento como um sucesso. O cientista de decisão está lá para ajudar a interpretar os dados, aplicar as melhores práticas para projetar experimentos e muito mais.

Regra curta: um cientista de decisão cria análises que produzem *recomendações*.

### **1.2.4 Trabalhos relacionados**

Embora as três áreas discutidas nas seções anteriores sejam os principais tipos de cargos da ciência de dados, você pode ver algumas outras funções distintas que estão fora dessas categorias. Listamos esses cargos a seguir, pois é bom saber que existem e pode ser preciso colaborar com colegas

nesses cargos. Então, se estiver interessado em um desses cargos, o material deste livro pode ser menos relevante para você.

## **Analista de inteligência de negócios**

Um *analista de inteligência de negócios* trabalha de forma similar a um analista, mas em geral usa menos estatística e programação. A ferramenta de escolha pode ser Excel em vez de Python, e é possível que nunca criem modelos estatísticos. Embora a função seja similar à de um analista, eles criam resultados menos sofisticados por causa das limitações de suas ferramentas e técnicas.

Se quiser fazer machine learning ou programação, ou aplicar métodos estatísticos, o cargo de analista de inteligência de negócios pode ser muito frustrante, pois não irá ajudá-lo a adquirir novas competências. Além disso, tais cargos via de regra pagam menos do que aqueles na ciência de dados, além de terem menos prestígio. Mas um cargo de analista de inteligência de negócios pode ser um bom ponto de partida para se tornar um cientista de dados, especialmente se não tiver trabalhado antes com dados em um ambiente de negócios. Se quiser começar como analista de inteligência de negócios e crescer para se tornar um cientista de dados, procure cargos nos quais possa aprender algumas competências que não tenha, como programação em R ou em Python.

## **Engenheiro de dados**

Um *engenheiro de dados* concentra-se em manter bancos de dados e garantir que as pessoas possam obter os dados de que precisam. Eles não fazem relatórios, análises ou desenvolvem modelos; em vez disso, mantêm os dados armazenados e formatados em bancos de dados bem estruturados para que outras pessoas possam fazer isso. Um engenheiro de dados pode ter a tarefa de manter todos os registros de clientes numa base de dados em grande escala na nuvem e adicionar novas tabelas a essa base de dados, conforme solicitado.

Os engenheiros de dados são bem diferentes dos cientistas de dados, além de serem mais raros e bastante procurados. Um engenheiro de dados pode ajudar a construir os componentes de back-end de dados do sistema de

experimento interno de uma empresa e atualizar o fluxo de processamento de dados quando os trabalhos começam a demorar muito. Outros engenheiros de dados desenvolvem e monitoram ambientes de lote e streaming, gerenciando dados da coleta ao processamento até armazenamento de dados.

Se estiver interessado em engenharia de dados, precisará de fortes competências em ciência da computação; muitos engenheiros de dados eram engenheiros de software.

## **Cientista de pesquisa**

Um *cientista de pesquisa* desenvolve e implementa novas ferramentas, algoritmos e metodologias, muitas vezes para serem usados por outros cientistas de dados dentro da empresa. Esses tipos de cargo quase sempre exigem doutorado, geralmente em ciência da computação, estatística, ciências sociais quantitativas ou em uma área relacionada. Os cientistas de pesquisa podem passar semanas pesquisando e experimentando métodos para aumentar o poder dos experimentos online, obtendo uma precisão 1% maior no reconhecimento de imagens em automóveis de autocondução ou construindo um novo algoritmo de deep learning (aprendizado profundo). Podem até mesmo investir tempo escrevendo artigos científicos, que raramente são usados na empresa, mas que ajudam a fortalecer o prestígio da empresa e (idealmente) promover avanços na área. Como esses cargos requerem conhecimentos muito específicos, não nos concentramos neles neste livro.

## **1.3 Como escolher sua direção**

No Capítulo 3, trataremos de algumas opções para obter competências em ciência de dados, os benefícios e os inconvenientes de cada opção, além de algumas sugestões para escolher. Neste ponto, é bom começar a refletir sobre a área da ciência de dados em que você quer se especializar. Onde você já tem experiência? Já vimos cientistas de dados que eram engenheiros, professores de psicologia, gerentes de marketing, estudantes de estatística e assistentes sociais. Muitas vezes, o conhecimento adquirido em outros empregos e áreas acadêmicas pode ajudá-lo a ser um cientista de

dados melhor. Se já estiver na ciência de dados, é bom refletir agora sobre em qual parte do triângulo se encontra. Você está feliz? Pretende mudar para um tipo diferente de trabalho na ciência de dados? A transição muitas vezes é possível.

**Vicki Boykis: qualquer um pode se tornar um cientista de dados?**



Com todo o otimismo (e possibilidade de grandes salários listados em notícias) em torno da ciência de dados, é fácil perceber por qual motivo apresenta oportunidades de carreira atrativas, especialmente pelo alcance e pelo escopo dos cargos da ciência de dados continuarem a expandir-se. Como novo candidato na área, porém, é importante ter uma visão realista e com as nuances para onde o mercado da ciência de dados está se dirigindo nos próximos anos e adaptar-se.

Há algumas tendências que afetam hoje a área da ciência de dados. Em primeiro lugar, a ciência de dados enquanto área já existe há dez anos e, como tal, passou pelas fases iniciais do ciclo da “modinha”: entusiasmo dos meios de comunicação social, adoção precoce e consolidação. Passou por exageros, falou-se nos meios de comunicação, foi adotada pelas empresas do Vale do Silício e muito mais. Estamos agora no ponto da adoção de um crescimento elevado em grandes empresas e da normalização de conjuntos de ferramentas de fluxo de trabalho da ciência de dados, como o Spark e o AutoML.

Em segundo lugar, como resultado, existe uma oferta excessiva de novos cientistas de dados, que vieram de bootcamps, de cursos de ciência de dados recém-instituídos em universidades ou de cursos online. O número de candidatos a um cargo da ciência de dados, em especial de nível básico, aumentou de 20 ou mais por vaga para 100 ou mais. Já não é incomum ver 500 currículos por vaga aberta.

Em terceiro lugar, a padronização das ferramentas e a pronta oferta de trabalho, bem como a procura por pessoas que tenham mais experiência no campo, significou uma mudança na forma como os cargos da ciência de dados são distribuídos e a criação de uma hierarquia de empregos e descrições da ciência de dados. Por exemplo, em algumas empresas, “cientista de dados” pode significar criar modelos, mas em outras significa, principalmente, executar análises de SQL, equivalente ao cargo de analista de dados.

Isso significa várias coisas para aqueles que procuram ingressar na ciência de dados como recém-chegados. Primeiro e mais importante, podem encontrar um mercado de trabalho extremamente competitivo e lotado, sobremaneira para aqueles que são novos no setor em geral (como recém-formados) ou aqueles que fazem a transição de outros

setores, competindo com milhares de candidatos que são exatamente como eles. Em segundo lugar, podem estar se candidatando a empregos que não são exatamente da ciência de dados, como é retratado em posts de blogs e na imprensa popular – apenas escrevendo e implementando algoritmos.

Tendo em conta essas tendências, é importante compreender que pode ser difícil diferenciar-se inicialmente de outros currículos para entrar na última rodada de entrevistas. Embora as estratégias que você lê neste livro possam parecer trabalhosas, elas o ajudarão a se destacar, o que é necessário neste novo e competitivo ambiente da ciência de dados.

## **1.4 Entrevista com Robert Chang, cientista de dados da Airbnb**

Robert Chang é cientista de dados do Airbnb, onde trabalha no produto Airbnb Plus. Trabalhou antes no Twitter, na equipe de Growth, fazendo análises de produtos e experimentos e criando modelos e pipelines de dados. É possível encontrar seus posts de blog em engenharia de dados, conselhos para novos e aspirantes a cientistas de dados e o trabalho dele no Airbnb e Twitter em <https://medium.com/@rchang>.

### **Qual foi seu primeiro passo na ciência de dados?**

Meu primeiro trabalho como cientista de dados foi no *The Washington Post*. Em 2012, eu estava pronto para deixar a universidade e entrar no setor, mas não sabia o que queria fazer. Esperava ser um cientista de visualização de dados, tendo ficado impressionado com o trabalho do *The New York Times*. Quando fui à feira de carreira da minha universidade e vi que o *The Washington Post* estava contratando, ingênuo como eu era, presumi que deveriam estar fazendo algo similar ao *The New York Times*. Eu me candidatei e consegui o emprego, não fazendo mais pesquisa prévia nenhuma.

Se perguntassem um exemplo de como não começar a carreira na ciência de dados, definitivamente seria o meu caso! Consegui o emprego esperando trabalhar com a visualização ou a modelação de dados, mas percebi muito rapidamente que meu trabalho era mais o de um engenheiro de dados. Muito do meu trabalho foi montar pipelines ETL (extract transform load), rodar scripts de SQL e tentar garantir que os relatórios fossem feitos para que pudéssemos relatar métricas de alto nível aos executivos. Foi muito doloroso na época; percebi que o que eu queria fazer não estava alinhado

com o que a empresa realmente necessitava e acabei saindo.

Nos meus anos seguintes no Twitter e no Airbnb, percebi que estava vendo a norma e não a exceção. Quando você está criando os recursos de dados, terá que construí-los camada a camada. Monica Rogati escreveu uma postagem famosa no blog sobre a hierarquia das necessidades da ciência de dados, que vai direto ao ponto (<http://mng.bz/ad0o>). Naquela altura, eu era novo demais para apreciar como o trabalho da ciência de dados era realizado de verdade e em tempo real.

## **O que as pessoas devem procurar em um trabalho de ciência de dados?**

Se estiver buscando um cargo na ciência de dados, deve se concentrar no estado da infraestrutura de dados da empresa. Se você trabalhar em uma empresa onde há apenas dados brutos que não são armazenados em um data warehouse, provavelmente levará meses ou, às vezes, até anos para chegar a um ponto em que você possa fazer análises, experimentos ou machine learning interessantes. Se isso não for algo que você espera fazer, terá um desalinhamento essencial entre o estágio da empresa e como quer contribuir com ela.

Para avaliar isso, dá para fazer perguntas do tipo: vocês têm uma equipe de infraestrutura de dados? Há quanto tempo eles trabalham aqui? O que é pilha de dados? Vocês têm uma equipe de engenharia de dados? Como eles trabalham com os cientistas de dados? Quando estão montando um novo produto, existe um processo para instrumentar logs, criar tabelas de dados e colocá-las no data warehouse? Se não houver, você fará parte da equipe responsável pela montagem disso, e pode considerar passar muito tempo nesta atividade.

A segunda coisa a olhar são as pessoas. Há três tipos de pessoas a quem prestar atenção. Presumindo que você não quer ser o primeiro cientista de dados, é melhor trabalhar em uma empresa de ciência de dados onde há um líder experiente. Um líder experiente sabe como construir e manter uma boa infraestrutura e fluxo de trabalho para que os cientistas de dados sejam produtivos. Em segundo lugar, procure um gerente que apoie o aprendizado contínuo. Por fim, é extremamente importante, sobretudo quando você é

novo no trabalho, trabalhar com uma liderança de tecnologia ou com um cientista de dados sênior que seja bastante prático. Para o dia a dia do seu trabalho, essa é a pessoa que mais o ajuda.

## **Quais competências são necessárias para ser um cientista de dados?**

Acho que depende do tipo de trabalho que está buscando e daquilo que o empregador define. As empresas de alto nível em geral exigem mais, às vezes até demais, porque há muitas pessoas tentando ingressar na empresa. Geralmente, estão procurando unicórnios – alguém que tenha competências na estruturação de dados com R ou Python, experiência na construção de pipelines ETL, engenharia de dados, design de experimentos e na construção e produção de modelos. É muita pressão para os candidatos! Embora todas essas competências possam ser aprendidas e sejam úteis para qualquer problema que possa surgir, não acho que sejam necessárias para entrar na ciência de dados.

Se você conhece R ou Python e um pouco de SQL, já está em uma posição bastante boa para ingressar na ciência de dados. Se puder planejar sua carreira de forma a aprender mais, é sempre útil, mas não acho que seja um requisito. É mais importante gostar de aprender. Se estiver tentando ser contratado por empresas de tecnologia de alto nível, você precisa de um pouco mais, mas é mais para causar boa impressão do que realmente precisar para o trabalho. É útil fazer a distinção entre as competências essenciais e necessárias para iniciar uma carreira na ciência de dados e outras que sejam boas para se ter caso queira entrar numa empresa competitiva e bem conhecida.

## **Resumo**

- As competências em ciências de dados variam entre pessoas e cargos. Embora algum conhecimento seja fundamental, os cientistas de dados não precisam ser especialistas em todas as áreas relevantes.
- Os cargos na ciência de dados têm diferentes áreas de foco: apresentar os dados certos e limpos aos stakeholders (análise), colocar os modelos

de machine learning em produção e utilizar os dados para tomar uma decisão (ciência da decisão).

## CAPÍTULO 2

# Empresas de ciência de dados

Este capítulo abrange:

- Os tipos de empresas que contratam cientistas de dados
- As vantagens e desvantagens de cada tipo de empresa
- Os recursos tecnológicos em vagas diferentes

Como discutido no Capítulo 1, a ciência de dados é uma área ampla e com muitas funções diferentes: cientista de pesquisa, engenheiro de machine learning, analista de inteligência de negócios e muito mais. Embora o trabalho como cientista de dados dependa da função, é igualmente influenciado pela empresa. Questões como empresa grande *versus* pequena, negócios voltados à tecnologia *versus* modelo tradicional, empresas jovens *versus* empresas consagradas podem influenciar o foco do projeto, a tecnologia de suporte e a cultura da equipe. Ao compreender alguns modelos de empresas, você estará mais bem preparado quando estiver buscando locais para trabalhar, seja para seu primeiro ou n-ésimo emprego em ciência de dados.

Este capítulo tem como objetivo dar uma ideia de como é trabalhar diariamente em algumas empresas. Vamos apresentar cinco empresas fictícias que contratam cientistas de dados. Nenhuma dessas empresas é real, mas todas se baseiam em pesquisas e em nossas próprias experiências de trabalho, além de ilustrarem princípios básicos que podem ser amplamente aplicados. Embora as empresas não sejam exatamente iguais, conhecer esses cinco modelos deve ajudá-lo a avaliar possíveis empregadores.

Embora esses estereótipos sejam baseados no que vimos como tendências no setor, certamente não são a norma. Você pode encontrar uma empresa que rompe totalmente o padrão do que dizemos aqui – ou uma equipe

específica na empresa que é diferente da própria empresa.

Apesar de as empresas apresentadas neste capítulo serem inventadas, todas as informações que você verá são de cientistas de dados reais que trabalham em empresas reais!

## **2.1 MTC (Massive Tech Company – Empresa de tecnologia em massa)**



- Similar a: Google, Facebook e Microsoft
- Idade da empresa: 20 anos
- Funcionários: 80 mil

A MTC é uma empresa de tecnologia com uma área de atuação enorme, vendendo serviços em nuvem, software de produtividade do consumidor, como editor de texto, hardware de servidor e inúmeras soluções empresariais pontuais. A empresa acumulou uma grande fortuna que utiliza para financiar projetos de pesquisa e desenvolvimento (P&D) incomuns, como scooters autodirigíveis e tecnologia de realidade virtual (RV). A equipe de P&D gera notícias, mas a maioria da equipe técnica é formada por engenheiros que fazem melhorias progressivas nos produtos existentes, adicionam mais funcionalidades, melhoram a interface do usuário e lançam novas versões.

### **2.1.1 Sua equipe: uma de muitas na MTC**

A MTC tem quase mil cientistas de dados espalhados pela empresa. Na sua maioria, esses cientistas de dados são agrupados em equipes, cada uma dando suporte a um produto ou divisão diferente, ou são colocados individualmente numa equipe que não é de ciência de dados a fim de dar suporte total. Existem, por exemplo, cientistas de dados de dispositivo de RV (VR headset) em uma equipe, cientistas de dados de marketing em uma segunda equipe e cientistas de dados de marketing de dispositivo de RV em



uma terceira, enquanto aquela da cadeia de fornecimento de dispositivo de RV também tem seu próprio cientista de dados.

Se fizesse parte de uma dessas equipes de ciência de dados, ao entrar na empresa, você teria sido escalado rapidamente. As grandes organizações contratam novas pessoas todos os dias; portanto, a empresa deve ter um procedimento padrão para alocar um computador portátil e dar acesso aos dados, bem como para treiná-lo na utilização de quaisquer ferramentas especiais. Na equipe, você seria encarregado de fazer ciência de dados para sua área de foco específica. Essa área poderia incluir a criação de relatórios e gráficos que os executivos poderiam usar para justificar o financiamento de projetos. Igualmente poderia construir modelos de machine learning que seriam entregues aos desenvolvedores de software para serem colocados em produção.

É provável que sua equipe seja grande e com pessoas experientes. Como a MTC é uma grande empresa de tecnologia bem-sucedida, ela tem uma área de atuação ampla para atrair candidatos muitos bons. Sua equipe será grande, de forma que as pessoas trabalham em tarefas quase não relacionadas; uma pessoa poderia fazer uma análise exploratória para um diretor em linguagem R, por exemplo, e outra poderia construir um modelo de machine learning em Python para uma equipe-irmã. O tamanho da equipe é uma bênção e uma maldição: você tem um grande corpo de cientistas de dados especializados para discutir ideias, mas a maioria deles provavelmente não tem familiaridade com as tarefas específicas nas quais está trabalhando. Além disso, existe uma hierarquia estabelecida na equipe. As pessoas mais experientes tendem a ser mais ouvidas porque têm mais experiência no campo e em lidar com diferentes departamentos na MTC.

O trabalho que sua equipe faz provavelmente traz um equilíbrio saudável para manter a empresa funcionando, como fazer relatórios mensais e providenciar atualizações trimestrais do modelo de machine learning, assim como fazer projetos novos e gerar uma previsão que nunca tenha sido feita antes. O gerente da equipe tem de equilibrar a enxurrada de solicitações de trabalho de ciência de dados de outras equipes, que ajudem essas equipes a curto prazo, com o desejo de fazer um trabalho inovador, mas não solicitado, que possa proporcionar benefícios a longo prazo. Com muito

dinheiro em caixa, a empresa pode dar-se ao luxo de realizar muito mais inovação e P&D do que outras empresas, um fato que se resume à vontade de testar novos projetos interessantes de ciência de dados.

### **2.1.2 A tecnologia: avançada, mas desarticulada na empresa**

A MTC é uma empresa de grande escala, mas, com organizações desse porte, é impossível evitar o uso de diferentes tipos de tecnologia. Um departamento pode armazenar dados de pedidos e clientes em um banco de dados do Microsoft SQL Server; outro pode manter registros no Apache Hive. Pior ainda, não só a tecnologia para armazenar dados pode estar desarticulada, como também os dados em si. Um departamento pode manter registros de clientes indexados por número de telefone; outro departamento, ainda, pode utilizar os endereços de email para indexar clientes.

A maioria das empresas do tamanho da MTC tem seus próprios recursos tecnológicos caseiros. Portanto, como cientista de dados na MTC, você precisa aprender maneiras específicas de consultar e usar dados que são particulares à MTC. Aprender essas ferramentas especializadas é ótimo para obter mais acesso dentro da MTC, mas o conhecimento que você adquire não pode ser transferido para outras empresas.

Como cientista de dados, você provavelmente usará várias ferramentas possíveis. Como a MTC é tão grande, ela tem muito suporte para as principais linguagens, como R e Python, que muitas pessoas usam. Algumas equipes também podem usar linguagens pagas, como SAS ou SPSS, mas essa situação é um pouco mais rara. Se quiser usar uma linguagem incomum que gosta, mas que outras poucas pessoas utilizam, como Haskell, você pode ou não ser liberado, dependendo do seu gerente.

Os recursos de machine learning variam dramaticamente dependendo da parte da empresa em que você se encontra. Algumas equipes usam microsserviços e contêineres para implantar modelos de forma eficiente, enquanto outras possuem sistemas de produção antiquados. A diversidade de recursos para a implantação de software dificulta a conexão com as APIs de outras equipes; não há um único local central para aprender e entender o que está acontecendo.

### **2.1.3 As vantagens e desvantagens da MTC**

Ser cientista de dados na MTC significa ter um trabalho impressionante numa empresa impressionante. Como a MTC é uma empresa de tecnologia, as pessoas sabem o que é um cientista de dados e o que de útil você pode fazer. O fato de conhecerem, no geral, sua função torna o trabalho muito mais fácil. O número elevado de cientistas de dados na empresa significa que há uma grande rede de suporte na qual você pode confiar se estiver tendo alguma dificuldade, bem como processos simples para aderir à empresa e obter acesso aos recursos necessários. Raramente você se encontrará estagnado e sozinho.

Ter vários cientistas de dados no seu entorno traz desvantagens também. Os recursos tecnológicos são complexos e difíceis de navegar, porque muitas pessoas a fizeram de muitas maneiras. Uma análise que pediram que você recriasse pode estar em uma linguagem que você não conhece, escrita por alguém que não está mais na empresa. Será mais difícil destacar-se e ser notado porque existem muitos outros cientistas de dados à sua volta, podendo ser difícil encontrar um projeto interessante para trabalhar, pois muitos dos projetos óbvios já foram iniciados por outras pessoas.

Como a MTC é uma empresa estabelecida, trabalhar nela dá uma sensação maior de segurança. Há sempre o risco de demissões, mas trabalhar em uma MTC não é como trabalhar em uma startup, onde o financiamento pode acabar a qualquer momento. Além disso, em grandes empresas, os gerentes preferem encontrar uma nova equipe para alguém trabalhar em vez de demiti-lo; demitir gera complicações legais que requerem um suporte de backup completo para a decisão da rescisão.

Profissionais que atuam em muitas funções especializadas na empresa podem se constituir em vantagem ou desvantagem da MTC. Engenheiros de dados, arquitetos de dados, cientistas de dados, pesquisadores de mercado e muitos outros profissionais desempenham diferentes funções relacionadas à ciência de dados, o que significa que terá muitas pessoas para você passar o trabalho. Você tem uma chance pequena de ser forçado a criar seu próprio banco de dados, por exemplo. Essa situação é excelente para passar adiante o trabalho que fica fora da sua especialidade, mas também significa que talvez você não aumente suas competências.

Outra desvantagem da MTC é a burocracia. Em uma grande empresa, obter aprovações para coisas como novas tecnologias, viagens a conferências e iniciar projetos pode exigir subir na cadeia de comando. Pior ainda, o projeto em que trabalha há anos pode ser cancelado porque dois executivos estão brigando e seu projeto é um dano colateral.

A MTC é uma empresa excelente para cientistas de dados que procuram ajudar a resolver grandes problemas utilizando técnicas de ponta – tanto cientistas de decisão que pretendem fazer análises quanto engenheiros de machine learning que pretendem construir e implementar modelos. Grandes empresas têm muitos problemas para resolver e um orçamento que permite tentar coisas novas. Você pode não ser capaz de tomar grandes decisões sozinho, mas você sabe que contribuiu.

A MTC é uma escolha ruim para um cientista de dados que quer ser tomador de decisões e receber o crédito. Uma grande empresa estabelece métodos, protocolos e estruturas que você precisa seguir.

## 2.2 HandbagLOVE: a varejista bem estabelecida

### HandbagLOVE

- Similar a: Payless, Bed Bath & Beyond e Best Buy
- Idade da empresa: 45 anos
- Tamanho: 15 mil funcionários (10 mil em lojas de varejo, 5 mil na área corporativa)

A HandbagLOVE é uma cadeia de varejista em 250 locais nos Estados Unidos, todas vendendo bolsas e clutches. A empresa existe há muito tempo e está repleta de especialistas em como criar uma loja e melhorar a experiência do cliente. A empresa é lenta na adoção de novas tecnologias, levando muito tempo para ter o primeiro site e primeiro aplicativo.

Recentemente, a HandbagLOVE teve uma queda nas vendas, já que a Amazon e outras varejistas online têm abocanhado sua quota de mercado. Sabendo que está ficando para trás, a HandbagLOVE tem procurado melhorar por meio da tecnologia, investindo em um aplicativo online e em uma capacitação da Amazon Alexa, tentando usar o valor de seus dados. A

HandbagLOVE possui analistas financeiros empregados há muitos anos que atuam no cálculo de estatísticas agregadas de alto nível sobre pedidos e clientes, mas apenas recentemente a empresa considerou a contratação de cientistas de dados para ajudá-los a entender melhor o comportamento dos clientes.

A recém-formada equipe de ciência de dados foi montada com base em analistas financeiros que antes elaboravam relatórios em Excel sobre métricas de desempenho para a empresa. Como a HandbagLOVE complementou essas pessoas com cientistas de dados treinados, a equipe começou a oferecer produtos mais sofisticados: previsões estatísticas mensais sobre o crescimento do cliente em linguagem R, dashboards interativos que permitem que os executivos entendam melhor as vendas e uma segmentação do cliente que reúne clientes em grupos úteis para marketing.

Embora a equipe tenha feito modelos de machine learning para alimentar novos relatórios e análises, a HandbagLOVE está longe de implantar modelos de machine learning em produção contínua. Quaisquer recomendações de produtos no site e no aplicativo deles são fornecidas por produtos de machine learning de terceiros, em vez de terem sido construídos na empresa. Fala-se na equipe de ciência de dados sobre a mudança dessa situação, mas ninguém sabe em quantos anos isso pode acontecer.

### **2.2.1 Sua equipe: um pequeno grupo com dificuldade para crescer**

A equipe conta basicamente com cientistas de dados que podem fazer relatórios em vez de serem capacitados em machine learning por ser uma área tão nova. Quando a equipe precisava de métodos modernos de machine learning e estatística, tinham de aprender por conta, já que ninguém sabia nada sobre eles. O autoaprendizado é ótimo, pois as pessoas começam a aprender novas técnicas que lhes são interessantes. A desvantagem é que alguns dos métodos técnicos utilizados podem ser ineficientes ou até mesmo errados porque não existem especialistas para conferir o trabalho.

A HandbagLOVE estabeleceu direções gerais para que os cientistas de

dados progridam para cargos que exijam mais experiência. Infelizmente, essas direções de carreira não são específicas para a ciência de dados; são objetivos de alto nível copiados e colados de outros cargos, como desenvolvimento de software, pois ninguém sabe realmente qual deve ser a métrica. Para progredir na carreira, você tem de convencer seu gerente de que está pronto e, com sorte, ele pode conseguir sua promoção. Do lado positivo, se a equipe acabar crescendo, você rapidamente se tornará um especialista na equipe.

Como a equipe de ciência de dados gera relatórios e modelos para departamentos de toda a empresa (como marketing, cadeia de suprimentos e atendimento ao cliente), a equipe de ciência de dados é bem conhecida. Esse fato concedeu à equipe grande respeito na empresa e, por sua vez, a equipe de ciência de dados tem bastante afinidade. A combinação do tamanho da equipe e do nível de influência dentro da empresa permite que os cientistas de dados tenham muito mais influência do que teriam em outras empresas. Não é incomum que alguém da equipe de ciência de dados se encontre com executivos de alto escalão e participe das reuniões.

### **2.2.2 Sua tecnologia: recursos legados que estão começando a mudar**

Uma frase comum que se ouve quando o assunto é tecnologia na HandbagLOVE é “sempre foi assim”. Os dados de pedidos e clientes são armazenados em um banco de dados Oracle diretamente conectado à tecnologia do caixa, o que não mudou em 20 anos. O sistema foi levado além dos próprios limites e sofreu muitas modificações, embora ainda funcione. Outros dados são coletados e armazenados na base de dados central, bem como: dados coletados no site, dados das chamadas de apoio ao cliente e dados de promoções e emails de marketing. Todos esses servidores ficam nas instalações (*no local*), não na nuvem, e uma equipe de TI os mantém.

Ao ter todos os dados armazenados em um servidor grande, você tem a liberdade de se conectar e juntar os dados como quiser, e, embora suas consultas às vezes levem muito tempo ou sobrecarreguem o sistema, dá para encontrar outra solução para obter algo utilizável. A grande maioria

das análises é feita no seu computador portátil. Se precisar de um computador mais potente para treinar um modelo, é bem difícil consegui-lo. A empresa não tem recursos de machine learning porque não a realiza internamente.

### **2.2.3 Vantagens e desvantagens da HandbagLOVE**

Na HandbagLOVE, você tem muita influência e capacidade para fazer o que acredita ser certo. Pode propor um modelo de valor vitalício para o cliente, construí-lo e usá-lo dentro da empresa sem ter que persuadir muitas pessoas para seguir com sua ideia. Essa liberdade, que se deve a uma combinação do tamanho da empresa e da novidade da ciência de dados, é muito gratificante; você tem capacidades incríveis para fazer o que pensa ser melhor. A desvantagem desse poder é que você não tem muitas pessoas para quem pedir ajuda. Você é responsável por encontrar uma maneira de fazer as coisas funcionarem ou lidar com as consequências quando as coisas não funcionam.

Os recursos tecnológicos são antiquados, e você terá que passar muito tempo encontrando soluções alternativas, o que não é um bom uso de tempo. Talvez queira usar uma tecnologia mais recente para armazenar dados ou executar modelos, mas não terá o suporte técnico para isso. Se não conseguir configurar uma nova tecnologia por si só, terá de seguir adiante sem ela.

O salário de um cientista de dados não será tão elevado como em empresas maiores, especialmente as de tecnologia. A HandbagLOVE simplesmente não tem o dinheiro disponível para pagar salários elevados para a ciência de dados. Além disso, a empresa não precisa dos melhores cientistas de dados – apenas pessoas que consigam fazer o básico. O salário não será horrível; certamente estará bem acima do que a maioria das pessoas na empresa ganha com anos de experiência semelhantes.

A HandbagLOVE é uma boa empresa para se trabalhar para cientistas de dados que estão entusiasmados por terem a liberdade de fazer o que pensam estar certo, mas talvez não estejam interessados em utilizar os métodos mais avançados. Se estiver confortável usando métodos estatísticos padrão e fazendo relatórios mais comuns, a HandbagLOVE pode ser um lugar

confortável para crescer na sua carreira. Se estiver realmente interessado em apenas utilizar métodos modernos de machine learning, você não encontrará muitos projetos na HandbagLOVE nem muitas pessoas que entendam do assunto.

## 2.3 Seg-Metra: a startup em fase inicial



- Similar a: milhares de startups falidas sobre as quais você não ouviu falar
- Idade da empresa: 3 anos
- Tamanho: 50 funcionários

A Seg-Metra é uma empresa jovem que vende um produto que ajuda as empresas de clientes a otimizar seus sites, personalizando-os para segmentos exclusivos de clientes. A Seg-Metra vende seu produto a empresas e não a consumidores. No início de sua breve história, a Seg-Metra teve alguns clientes de grande nome para começar a usar a ferramenta, o que a auxiliou na obtenção de mais financiamento de capitalistas de risco. Agora, com milhões de dólares na mão, a empresa está buscando crescimento rápido e melhora do produto.

A maior melhoria que os fundadores têm vendido aos investidores é agregar métodos básicos de machine learning ao produto. Essa melhoria foi vendida aos investidores como “IA (inteligência artificial) de ponta”. Com esse novo financiamento, os fundadores estão buscando engenheiros de machine learning para construir o que foi vendido. Eles também precisam de cientistas de decisão para começar a reportar sobre o uso da ferramenta, permitindo que a empresa compreenda mais claramente as melhorias que deve fazer no produto.

### 2.3.1 Sua equipe (que equipe?)



Dependendo de quando um cientista de dados for contratado, ele pode muito bem ser o primeiro em sua empresa. Se não for o primeiro, estará entre as contratações iniciais de ciência de dados e, provavelmente, reportará a quem foi contratado primeiro. Devido à novidade da equipe, haverá poucos ou nenhum protocolo – nenhuma linguagem de programação estabelecida, nem melhores práticas, maneiras de armazenar código ou reuniões formais.

Qualquer direção virá da contratação desse primeiro cientista de dados. A cultura da equipe provavelmente será definida pela benevolência dele. Se essa pessoa estiver aberta à discussão em grupo e a confiar nos outros membros do grupo, a equipe de ciência de dados como um todo decidirá assuntos como, por exemplo, qual linguagem usar. Se a pessoa for controladora e não estiver aberta a ouvir, ela tomará as decisões por si mesma.

Um ambiente não estruturado pode criar uma camaradagem enorme. Toda a equipe de ciência de dados trabalha arduamente, passa por dificuldades para fazer novas tecnologias, métodos e ferramentas funcionarem, podendo formar laços fortes e gerar amizades. Já aqueles que detêm poder poderiam infligir imensos abusos emocionais aos subordinados e, como a empresa é pequena, há pouca responsabilidade. Independentemente de como o crescimento da Seg-Metra seguir, os cientistas de dados nessa empresa de fase inicial terão uma trajetória com muitos obstáculos.

O trabalho da equipe pode ser fascinante ou frustrante, dependendo do dia. Muitas vezes, os cientistas de dados estão realizando análises pela primeira vez, como, por exemplo, fazer a primeira tentativa de usar dados de compra de clientes para segmentar clientes ou implantar a primeira rede neural na produção. Essas primeiras análises e tarefas de engenharia são empolgantes porque se constituem em território desconhecido dentro da empresa, e os cientistas de dados são os pioneiros. Em outros dias, o trabalho pode ser difícil, como quando uma demonstração tem de estar pronta para um investidor, mas o modelo ainda não está convergindo o dia anterior. Mesmo que a empresa tenha dados, é possível que a infraestrutura seja tão desorganizada que os dados não possam ser utilizados. Embora o trabalho seja caótico, todas essas tarefas significam que os cientistas de

dados aprendem muitas habilidades muito rapidamente trabalhando na Seg-Metra.

### **2.3.2 A tecnologia: tecnologia de ponta**

Por ser uma empresa jovem, a Seg-Metra não está condicionada a ter de manter uma antiga tecnologia legada. A Seg-Metra também quer impressionar seus investidores, o que é muito mais fácil de fazer quando os recursos tecnológicos são significativos. Assim, a Seg-Metra conta com os mais recentes e melhores métodos de desenvolvimento de software, armazenamento e coletando dados, além de fazer a análise e a geração de relatórios. Os dados são armazenados em uma variedade de tecnologias modernas na nuvem, e nada é feito no local. Os cientistas de dados se conectam diretamente a esses bancos de dados e criam modelos de rede neural de machine learning em grandes instâncias de máquinas virtuais da Amazon Web Services (AWS) com processamento de GPU. Esses modelos são implantados por meio de métodos modernos de engenharia de software.

À primeira vista, os recursos são certamente impressionantes. A empresa é tão jovem e está crescendo com tanta rapidez que as questões surgem continuamente com as diferentes tecnologias trabalhando em conjunto. Quando, subitamente, os cientistas de dados notam a falta de dados no armazenamento em nuvem, eles precisam esperar que o engenheiro de dados com excesso de trabalho conserte (e isso se tiverem a sorte de ter um engenheiro de dados). Seria ótimo se a Seg-Metra tivesse uma equipe dedicada de operações de desenvolvimento (DevOps) para ajudar a manter tudo funcionando, mas, até agora, o orçamento foi investido em outro lugar. Além disso, a tecnologia foi instalada tão rapidamente que, embora a empresa seja jovem, seria difícil monitorar tudo isso.

### **2.3.3 As vantagens e desvantagens da Seg-Metra**

Como uma startup em crescimento, a Seg-Metra tem muito apelo. O crescimento da empresa está oferecendo todo tipo de trabalho interessante em ciência de dados e um ambiente no qual cientistas de dados são forçados a aprender rapidamente. Esse tipo de cargo pode ensinar competências que aceleram o crescimento de uma carreira na área da ciência de dados –

competências como trabalhar com prazos com restrições limitadas, comunicar-se de maneira eficaz com pessoas de outras áreas e saber quando deve prosseguir com um projeto ou não. Especialmente no início de uma carreira, desenvolver essas competências pode torná-lo muito mais atraente como funcionário do que aqueles que têm trabalhado apenas em empresas maiores.

Outra vantagem que a Seg-Metra oferece é que você começa a trabalhar com as tecnologias mais recentes. A utilização da tecnologia mais recente deve tornar seu trabalho mais agradável: presumivelmente, as novas tecnologias que surgem são melhores do que as antigas. Ao aprender a última tecnologia, você também deve ficar com um currículo mais rico para futuros empregos. As empresas que pretendem utilizar tecnologia mais recente vão querer a sua ajuda.

Embora o pagamento não seja tão competitivo como em empresas maiores, especialmente aquelas de tecnologia, o trabalho fornece opções de ações, com o potencial de serem extremamente valiosas. Se a empresa eventualmente colocar suas ações em oferta pública ou for vendida, aquelas opções poderiam valer centenas de milhares de dólares ou mais. Infelizmente, a probabilidade de isso acontecer é bastante variável. Ou seja, esse fato só é uma vantagem se você gostar de apostas.

Uma desvantagem de trabalhar na Seg-Metra é que você tem que se empenhar muito. Trabalhar de 50 a 60 horas semanais não é incomum, e a empresa espera que todos contribuam da melhor maneira possível. Aos olhos da empresa, se todos não estiverem trabalhando juntos, ela não será bem-sucedida, então você realmente será o único funcionário a tirar férias em um ano? Esse ambiente pode ser extremamente tóxico, típico para abuso e muito desgastante.

A empresa é volátil, baseando-se em encontrar novos clientes e na ajuda de investidores para seguir funcionando, dando à Seg-Metra a desvantagem de baixa segurança de trabalho. É possível que, em qualquer ano, a empresa possa decidir demitir pessoas ou parar completamente de funcionar. Essas alterações podem ocorrer sem aviso prévio. A insegurança no emprego é sobremaneira difícil para as pessoas que têm famílias, o que faz com que o perfil de seus funcionários seja mais jovem. Uma força de trabalho jovem

também pode ser uma desvantagem se você quiser trabalhar com uma equipe mais diversificada e experiente.

No geral, trabalhar na Seg-Metra oferece uma grande oportunidade para atuar com tecnologia interessante, aprender muito rapidamente e ter uma pequena chance de ganhar uma tonelada de dinheiro. Entretanto, requer uma quantidade imensa de trabalho e um ambiente potencialmente tóxico. Por isso, essa empresa é melhor para cientistas de dados que procuram obter experiência e, em seguida, ir adiante.

**Rodrigo Fuentealba Cartes, cientista-chefe de dados em uma pequena empresa de consultoria governamental**

A empresa em que trabalho fornece análises, ciência de dados e soluções mobile para instituições governamentais, forças armadas e policiais e para alguns clientes privados. Sou o cientista-chefe de dados e o único responsável por projetos de ciência de dados. Não temos engenheiros de dados, técnicos de processamento de dados ou qualquer outra função de ciência de dados, porque o departamento é relativamente novo. Em vez disso, temos administradores de banco de dados, desenvolvedores de software e integradores de sistemas. Também desempenho a função de arquiteto de sistema/software e desenvolvedor de software livre. Pode parecer estranho, mas trabalho sob pressão e tudo funciona surpreendentemente bem.

Uma história estranha do meu trabalho: estava atuando em um projeto que envolvia o uso de informações históricas de muitas variáveis ambientais, como condições climáticas diárias. Faltaram dados críticos, pois uma área de estudo não tinha estações meteorológicas instaladas. O projeto estava em perigo, e o cliente decidiu encerrá-lo em uma semana caso não pudessem encontrar as informações.

Decidi voar para a área e entrevistar alguns pescadores, e perguntei-lhes como sabiam que era seguro navegar. Responderam que geralmente enviavam um navio que transmitia as condições meteorológicas por rádio. Visitei uma estação de rádio, e eles tinham transcrições manuscritas de comunicações desde 1974. Implementei um algoritmo que conseguia reconhecer notas manuscritas e extraí informações significativas e, em seguida, implementei um pipeline de processamento de linguagem natural que pudesse analisar as strings. Graças à saída a campo e encontrar esses dados incomuns, o projeto foi salvo.

**Gustavo Coelho, cientista-chefe de dados em uma pequena startup**

Estou trabalhando nos últimos 11 meses em uma startup relativamente nova, que se concentra na aplicação de IA à gestão de RH. Prevemos o desempenho futuro dos candidatos ou a probabilidade de serem contratados por uma determinada empresa. Essas previsões destinam-se a ajudar na aceleração do processo de contratação. Dependemos fortemente da atenuação de viés em nossos modelos. É uma pequena empresa: temos 11 pessoas e a equipe de ciência dos dados é composta por cinco delas, incluindo eu. Toda a empresa se dedica a auxiliar a equipe de ciência de dados a fornecer os modelos treinados para a produção.

Trabalhar em uma startup pequena me dá a oportunidade de aprender novos conceitos e aplicá-los todos os dias. Adoro pensar na melhor forma de configurar nossos processos de ciência de dados para que possamos escalar e dar mais liberdade aos nossos cientistas de dados para se concentrarem na ciência de dados. O RH não é um campo com conhecimentos técnicos, então, investe-se mais de metade da duração do projeto explicando a solução aos nossos clientes e os ajudando a sentirem-se confortáveis com os novos conceitos. Quando finalmente obtemos a aprovação, muito tempo também é investido na coordenação com o departamento de TI do cliente para integrar-se ao nosso pipeline de dados.

## **2.4 Videory: a bem-sucedida startup de tecnologia de estágio avançado**



- Semelhante a: Lyft, Twitter e Airbnb
- Idade da empresa: 8 anos
- Tamanho: 2 mil pessoas

A Videory é uma bem-sucedida startup de tecnologia de estágio avançado que possui uma rede social com base em vídeo. Os usuários podem carregar vídeos de 20 segundos e compartilhá-los com o público. A empresa acabou

de colocar suas ações como oferta pública, e todos estão em êxtase. A Videory não está perto do tamanho da MTC, mas está se saindo bem como uma rede social e aumentando todos os anos a base de clientes. É especialista em dados e provavelmente tem analistas ou cientistas de dados há alguns anos ou até mesmo desde o seu início. Os cientistas de dados da equipe estão muito ocupados fazendo análises e relatórios para dar suporte ao negócio, bem como criando modelos de machine learning para ajudar a combinar pessoas a artistas para trabalho comissionado.

### **2.4.1 A equipe: especializada, mas com espaço para mover-se**

A Videory ainda está no nível em que é possível reunir todos os cientistas de dados em uma grande sala de conferências. Dada a dimensão da empresa, a equipe pode ser organizada em um modelo centralizado. Cada pessoa da ciência de dados reporta-se a um gerente de ciência de dados, e todos estão em um único departamento grande da empresa. A equipe central de ciência de dados auxilia outros grupos em toda a empresa, mas a equipe define suas próprias prioridades. Alguns cientistas de dados estão até mesmo trabalhando em projetos internos de pesquisa acadêmica de longo prazo cujos benefícios não são imediatos.

A equipe de ciência de dados da Videory tem focos mais definidos, dada a dimensão da empresa. Há também algumas delineações entre as pessoas que fazem a parte pesada do machine learning, estatística ou análises. A Videory é pequena o bastante para poder fazer alternações entre esses grupos ao longo do tempo. Os cientistas de dados geralmente têm alguma interação – como sessões de treinamento, reuniões mensais e um canal de Slack compartilhado –, o que não se encontraria em empresas como a MTC, que são muito grandes para todos compartilharem ao mesmo tempo. É provável que as subequipes usem diferentes ferramentas e que um grupo de pessoas com doutorado publique documentos acadêmicos e ainda faça trabalhos teóricos.

### **2.4.2 A tecnologia: evitando se complicar com código legado**

A Videory tem um grande número de código e tecnologia legados, além de, provavelmente, pelo menos algumas ferramentas que foram desenvolvidas



internamente. É provável que a empresa esteja tentando acompanhar os desenvolvimentos tecnológicos e pretenda mudar para um novo sistema ou complementar os existentes com novas tecnologias. Como na maioria das empresas, um cientista de dados indubitavelmente consultará um banco de dados SQL para obter dados. É provável que a empresa tenha algumas ferramentas de inteligência de negócios também, pois há muitos consumidores que não são da ciência de dados.

Como cientista de dados na Videory, você com certeza aprenderá algo novo. Todas essas empresas têm grandes volumes de dados e sistemas para lidar com isso. SQL não será suficiente; a empresa precisa processar bilhões de eventos todos os meses. No entanto, você pode tentar o Hadoop ou o Spark quando for necessário retirar alguns dados personalizados que não estejam armazenados no banco de dados SQL.

A ciência de dados é tipicamente feita em R ou Python, com muitos especialistas disponíveis para oferecer assistência diante de dificuldades que venham a surgir. O machine learning é implementado por meio de práticas modernas de desenvolvimento de software, como a utilização de microsserviços. Como a empresa é bem conhecida por ser uma startup bem-sucedida, muitas pessoas talentosas trabalham lá, usando abordagens de ponta.

### **2.4.3 As vantagens e as desvantagens da Videory**

A Videory pode ser de bom tamanho para os cientistas de dados. Embora existam outros dados suficientes para fornecer orientação e suporte, a equipe ainda é pequena o bastante para que todos se conheçam. A ciência de dados é reconhecida na empresa como sendo importante, o que significa que seu trabalho pode obter reconhecimento de vice-presidentes e talvez até mesmo do conjunto C (CEO, CTO etc.). Haverá engenheiros de dados para dar suporte ao seu trabalho. Os pipelines de dados podem ficar lentos às vezes ou até mesmo quebrarem, mas o funcionário não será responsável pela reparação dos mesmos.

Em uma empresa com mais de mil funcionários, você terá de lidar com questões políticas inevitáveis. É possível que seja pressionado a gerar números que correspondam ao que as pessoas querem ouvir (e que possam

dizer aos respectivos chefes para conseguir um bônus) ou enfrentar expectativas irrealistas sobre quão rápido algo pode ser desenvolvido. Você também pode acabar trabalhando em coisas que a empresa realmente não precisa, porque seu gerente pediu. Por vezes, acabará sentindo que não recebeu orientação ou que desperdiçou seu tempo. Embora não mude tanto quanto em uma startup de fase inicial, a empresa continuará mudando muito, e o que é uma prioridade em um trimestre pode ser totalmente ignorada no próximo.

Embora outros cientistas de dados da Videory conheçam mais do que você sobre a maioria dos tópicos de ciência de dados, você pode rapidamente se tornar especialista em algo específico, como a análise de séries temporais. Essa situação pode ser ótima se gosta de ensinar outras pessoas, em especial se seu trabalho permite tomar tempo para aprender mais sobre uma área particular, pela leitura de artigos ou por meio de cursos. Pode ser difícil quando você sente que ninguém consegue conferir seu trabalho ou motivá-lo a aprender coisas novas. Você sempre terá mais a aprender, mas o que aprende pode não estar na área na qual quer se concentrar.

No geral, a Videory fornece uma boa combinação de alguns dos benefícios dos outros modelos. É suficientemente grande para ter pessoas à sua volta que ajudem e prestem assistência quando necessário, mas não tão grande que as solicitações fiquem presas na loucura burocrática ou que os departamentos se sobreponham no escopo. Os cientistas de dados que trabalham na empresa têm muitas oportunidades de aprender, mas, devido à especialização das funções, não têm a oportunidade de experimentar tudo. Essa empresa é um ótimo lugar para os cientistas de dados que estão procurando uma aposta segura e que oferece chances de crescer, mas não um grande número de chances.

**Emily Bartha, a primeira cientista de dados em uma startup de médio porte**

Trabalho em uma startup de médio porte que tem um produto focado em seguros. Sendo a primeira cientista de dados, posso ajudar a definir nossa estratégia em torno da utilização de dados e da introdução de machine learning em nosso produto. Fico na equipe de dados da empresa, então trabalho muito de perto de engenheiros de dados e com o gerente de produtos de dados.

Meu dia de trabalho começa com uma reunião com a equipe de dados. Falamos sobre o que temos planejado para o dia e sobre o que está parado ou pendente. Passo muito tempo destrinchando dados: visualizando, criando relatórios e investigando problemas de qualidade ou esquisitices nos dados. Também dedico muito tempo à documentação. Quando programo, uso o GitHub, como o resto da equipe de engenharia, e tenho pessoas da equipe que revisam meu código (e eu reviso o delas). Além disso, participo de reuniões ou trabalho em colaboração com a minha equipe.

Como já atuei também em empresas maiores, amo trabalhar em uma empresa pequena! Aqui, há muita liberdade para tomar iniciativa. Se você tem uma ideia e quer trabalhar para torná-la realidade, ninguém cruzará seu caminho. Procure uma empresa que já tenha investido em engenharia de dados. Quando cheguei, já havia vários engenheiros de dados e uma estratégia para instrumentação, coleta de dados e armazenamento. Quando você trabalha em uma empresa pequena, as coisas estão em constante modificação e as prioridades mudam, então é importante adaptar-se com facilidade. As pessoas que gostam de mergulhar profundamente em um projeto e trabalhar nele por meses podem não gostar de trabalhar em uma startup, uma vez que muitas vezes requer o desenvolvimento de soluções que sejam boas o suficiente e passar para o próximo projeto.

## 2.5 Global Aerospace Dynamics: uma gigante do governo

### GLOBAL AEROSPACE DYNAMICS

- Similar a: Boeing, Raytheon e Lockheed Martin
- Idade da empresa: 50 anos
- Tamanho: 150 mil pessoas

A Global Aerospace Dynamics (GAD) é uma empresa enorme e rica, que traz dezenas de bilhões de dólares em receita todos os anos graças a vários contratos governamentais. A empresa desenvolve tudo, desde caças e mísseis a sistemas inteligentes de semáforos. Ela está espalhada por todo o país, em várias divisões, a maioria das quais não se comunica com as outras. A GAD existe há décadas, e muitos de seus funcionários estão lá há décadas também.

A GAD foi lenta na aceitação de ciência de dados. A maioria das divisões de engenharia tem coletado dados, mas com dificuldade em compreender como podem utilizá-los nos processos existentes, os quais são muito regimentados. Devido à natureza do trabalho, o código precisa ser extremamente improvável de ter bugs e testado com rigor. Portanto, a ideia

de implementar um modelo de machine learning, que tem previsibilidade limitada, é, na melhor das hipóteses, arriscada. De um modo geral, o ritmo de trabalho na empresa é lento; o lema do mundo da tecnologia “move fast and break things” (mova-se rapidamente e quebre as coisas) é o oposto da mentalidade da GAD.

Com a quantidade de artigos sobre inteligência artificial, a ascensão do machine learning e a necessidade de usar dados para transformar um negócio, os executivos da GAD estão prontos para dar início à contratação de cientistas de dados, os quais estão surgindo em equipes em toda a empresa, realizando tarefas como analisar dados de engenharia para gerar relatórios melhores, criar modelos de machine learning para colocar em produtos e trabalhar como provedores de serviços para ajudar os clientes da GAD a solucionarem problemas.

### **2.5.1 A equipe: um cientista de dados em um mar de engenheiros**

Embora as funções dependam de onde se encontram na GAD e em que projeto estão trabalhando, o cientista de dados médio é o único em uma equipe de engenheiros. Na melhor das hipóteses, podem existir dois ou três cientistas de dados na equipe, cuja função é apoiar os engenheiros com análise, construção de modelos e entrega de produtos. A maioria dos engenheiros da equipe tem apenas um conhecimento muito vago de ciência de dados; eles se lembram de algo da faculdade, mas não conhecem os princípios básicos da coleta de dados ou da engenharia de recursos, as dificuldades de validar um modelo ou como os modelos são implantados. Você terá poucos recursos para ajudá-lo quando as coisas não correrem bem, mas, uma vez que poucas pessoas entendem seu trabalho, mais ninguém pode notar que as coisas estão indo mal.

Muitos dos engenheiros da equipe estão na empresa há dez anos ou mais, então, eles têm muito conhecimento institucional. Será também mais provável que tenham a mentalidade de “estamos fazendo as coisas desta forma desde que estou aqui, por que, então, devemos mudar?”. Essa atitude tornará mais difícil a implementação das ideias propostas pelos cientistas de dados. A natureza mais lenta da indústria da defesa significa que as pessoas

tendem a trabalhar menos do que em outros locais; as pessoas batem o ponto 40 horas por semana, mas casualmente trabalhar menos não é incomum. Em outras empresas, você pode ficar sobrecarregado, já na GAD o estresse vem de não ter bastante trabalho a ser feito e entediar-se.

As promoções e os aumentos são extremamente fechados, pois os gerentes têm de seguir regras para reduzir o viés (e, portanto, menos probabilidade de a GAD ser processada) e também porque é assim que as coisas são feitas há décadas. Receber aumentos e promoções tem a ver em grande parte com há quantos anos você trabalha na empresa. Trabalhar bastante pode fazer sua promoção chegar um ano mais cedo ou fazer com que você ganhe um bônus mais alto, mas há poucas chances de um cientista de dados júnior chegar rapidamente a ser um cientista-chefe de dados. O outro lado é que os funcionários raramente são demitidos.

### **2.5.2 A tecnologia: antiga, endurecida e com bloqueio de segurança**

Embora os recursos tecnológicos variem muito entre os grupos na GAD, tudo tende a ser relativamente antigo, rodando localmente em vez de na nuvem e coberto por protocolos de segurança. Como os dados envolvidos abrangem tópicos como desempenho de aviões de caça, é fundamental que os dados não vazem. Além disso, a empresa precisa de responsabilidade legal para qualquer tecnologia que utilize caso algo venha a correr mal, razão pela qual o código aberto tende a não ser bem recebido. Embora o Microsoft SQL Server seja mais caro do que o PostGRES SQL, por exemplo, a GAD tem o prazer de pagar à Microsoft o dinheiro extra, sabendo que, se houver um bug de segurança, pode chamar a Microsoft para lidar com ele.

Na prática, essa configuração se assemelha com dados sendo armazenados em bancos de dados do SQL Server, executados por uma equipe de TI extremamente mesquinha sobre quem tem acesso ao quê. Os cientistas de dados têm permissão para acessar os dados, mas devem executar Python em servidores especiais com acesso limitado à internet a fim de que nenhuma biblioteca envie secretamente dados para países estrangeiros. Se os cientistas de dados quiserem usar software de código aberto especial, há

poucas chances de que a TI e a segurança o aprovem, o que dificulta muito o trabalho dos mesmos.

Se o código precisar ser implantado em sistemas de produção, ele tende a ser colocado de maneiras tradicionais. A GAD está apenas começando a adotar métodos modernos para colocar o código de machine learning em produção.

### **2.5.3 As vantagens e desvantagens da GAD**

As vantagens de trabalhar na GAD é que se trata de empregos na ciência de dados confortáveis e seguros. O ritmo menos rigoroso do trabalho significa que é mais provável que você tenha energia quando chegar em casa à noite. Muitas vezes, você terá tempo livre quando estiver trabalhando, podendo ler blogs e artigos de ciência de dados sem que ninguém desaprove. O fato de outras poucas pessoas conhecerem os conceitos básicos de ciência de dados significa que terá menos pessoas o questionando. E como a GAD é uma empresa enorme que está preocupada com responsabilidades legais, você realmente teria que ficar aquém do desempenho para ser demitido.

A desvantagem de trabalhar na GAD é que o funcionário tem menos probabilidade de aprender novas habilidades do que teria em outras empresas. Provavelmente você será designado a um único projeto por anos, de modo que as tecnologias e as ferramentas usadas rapidamente se tornarão comuns. Pior ainda, as competências que vier a aprender serão para tecnologias desatualizadas que não são transferíveis para outras instituições. E, embora não seja demitido com facilidade, também não será facilmente promovido.

A GAD é um ótimo lugar para trabalhar se encontrar uma equipe fazendo projetos que você acha interessante e se não quer que o trabalho seja sua vida. Muitas pessoas trabalham para a GAD durante décadas porque é confortável e estão felizes por estarem confortáveis. Mas se você precisa de desafios, a GAD pode não ser uma boa opção.

**Nathan Moore, gerente de análise de dados em uma empresa de serviços públicos**

A empresa em que trabalho fornece e vende energia para centenas de milhares de pessoas, sendo parcialmente de propriedade do governo. A própria empresa tem cerca de mil funcionários espalhados por muitas funções diferentes. Meu trabalho envolve investigar e prototipar novas fontes de dados e atuar com os especialistas de banco de dados para limpar e documentar fontes de dados atuais. Temos muitos sistemas legados e novas iniciativas acontecendo, então, há sempre algo a se fazer.

No momento, o dia envolve reuniões, revisão de especificações para ETL, teste de uma nova técnica de machine learning que encontrei no Twitter, dar feedback sobre relatórios, aprender a usar JIRA e Confluence, e responder a muitos emails. Antigamente, envolvia-me no desenvolvimento e na avaliação de modelos, na análise de dados quando algum processamento durante a noite falhava e na apresentação de propostas ao governo sobre uma análise do setor em toda a indústria.



Somos grandes o suficiente para termos uma boa equipe de analistas para trabalhar em diversos projetos, desde relatórios diários até grandes projetos de segmentação de clientes. Tive muitas oportunidades na empresa e trabalho aqui há 11 anos. Mas como temos bilhões de dólares de ativos, a aversão ao risco é alta e o ritmo de mudança, um pouco lento. Temos um departamento de TI grande o suficiente para dar suporte às funções diárias, mas qualquer projeto significativo, como a atualização de sistemas, quer dizer que os recursos são escassos para qualquer melhoria não prioritária. Tudo precisa ser justificado e o orçamento reservado, além de muita política para se navegar.

## 2.6 Reunindo tudo

Quando você estiver buscando empresas para trabalhar, descobrirá que muitas delas são semelhantes de várias maneiras. À medida que você passa por candidaturas e entrevistas de emprego, pode ser útil tentar entender os pontos fortes e fracos do trabalho nessas empresas (Tabela 2.1).

*Tabela 2.1 – Um resumo das empresas que contratam cientistas de dados*

Critérios	MTC	HandbagLOVE	Seg-Metra	Videory	GAD
	Tecnologia em massa	Varejista	Startup	Empresa de tecnologia de médio porte	Defesa
Burocracia	Muita	Pouca	Nenhuma	Alguma	Muita
Recursos tecnológicos	Complexos	Antigos	Frágeis	Mistos	Antigos
Liberdade	Pouca	Muita	Bastante	Muita	Nenhuma
Salário	Incrível	Decente	Baixo	Ótimo	Decente
Segurança de trabalho	Ótima	Decente	Baixa	Decente	Ótima
Oportunidade de aprendizado	Muita	Alguma	Muita	Muita	Pouca

## 2.7 Entrevista com Randy Au, pesquisador quantitativo de experiência do usuário na Google

Randy Au trabalha na equipe do Google Cloud. Por ter trabalhado na ciência de dados com foco no comportamento humano há mais de uma década, ele escreve sobre como pensar a respeito de atuar em startups e diferentes tipos de empresas em [https://medium.com/@randy\\_au](https://medium.com/@randy_au).

## **Existem grandes diferenças entre empresas grandes e pequenas?**

Sim. Normalmente, mais em termos organizacionais e estruturais. Há pontos em uma empresa nos quais a cultura muda por causa do tamanho. Em uma startup de 10 pessoas, todos fazem tudo porque todos desempenham todas as funções. Entretanto, com cerca de 20 pessoas, as coisas começam a se especializar. Começa a ter equipes de três ou quatro pessoas dedicadas a tarefas específicas. As pessoas podem pensar mais sobre certos assuntos, e você não precisa aprender tudo sobre uma empresa. Com cerca de 80 a 100 pessoas, as equipes já não aumentam mais. Então, há mais processos para cada projeto. Você não conhece mais todas as pessoas na empresa. Não se sabe o que todos estão fazendo e, por isso, há muito mais trabalho para chegar a um entendimento comum. Além disso, depois de cerca de 150 a 200 pessoas, é impossível saber o que se passa na empresa, e é por isso que a burocracia tem de existir. Então, você vai à Google, que tem 100 mil pessoas. Lá, você não tem nenhuma ideia do que a maioria está fazendo.

Quanto menor for a empresa, maior será a probabilidade de interagir com todos. Em uma empresa de 40 pessoas, o CEO se sentaria na minha mesa, pois ambos estamos explorando um conjunto de dados juntos. Nunca acontecerá algo do tipo na Google. Mas você se sentiria bem com uma situação que acontece em muitas startups, em que você está construindo um carro de F1 e o está conduzindo ao mesmo tempo, e todos estão discutindo se você deve ficar com o volante? Quando você é a pessoa responsável pelos dados de uma empresa pequena, os métodos não são realmente importantes; você está apenas tentando obter todos os dados e algumas ideias sobre ele. Está tudo bem em não ser tão rigoroso de forma que possa tomar decisões mais rapidamente.

## **Há diferenças baseadas no setor da empresa?**

Alguns setores historicamente têm funcionários da matemática ou de dados. Uma empresa de seguros tem atuários, por exemplo. Essas pessoas estão aí há cem anos e realmente conhecem as estatísticas. Se uma empresa de seguros trouxer cientistas de dados, eles chegam com uma visão um pouco

diferente. Eles já têm essa estrutura incorporada para estatísticos extremamente talentosos. Eles preencherão alguma outra lacuna: haverá uma lacuna no big data ou na otimização do site ou de alguma outra coisa.

O setor financeiro também tem uma longa tradição de ter pessoas de exatas. Lembro-me de não ter passado em uma entrevista de uma vaga dessa área porque fizeram um teste de programação. Como cientista de dados, só garanto que meu código funcione e que dê a resposta correta; não penso muito sobre o desempenho até que vire um problema. O teste de programação avaliava o desempenho e descontava pontos por não ter um desempenho automático. Pensei comigo: “Vocês são do financeiro. Qual é o sentido?”.

Penso que, se falarmos com todos os que estão trabalhando na ciência de dados, a grande maioria, mas silenciosa, é de quem está fazendo esse tipo de trabalho pesado que não é nada atrativo. Recebi uma quantidade ridícula de respostas ao artigo que escrevi sobre ciência de dados em startups, com leitores dizendo: “Aham, essa é a minha vida”. Não é o que as pessoas dizem quando falam sobre ciência de dados. Não é o chamariz “aqui está um novo algoritmo brilhante que apliquei conforme o artigo da arXiv”. Acho que nunca apliquei algo de um artigo da arXiv nos meus 12 anos de trabalho. Ainda estou usando regressão porque ela realmente funciona! Penso que é essa a realidade.

Você terá de limpar dados; não acredito que haja alguém mesmo nos Facebooks e Googles da vida que não precise limpar dados. Talvez seja um pouco mais fácil limpar os dados porque há uma estrutura em torno deles. Mas, de qualquer forma, você terá de limpar os dados. É um fato da vida.

## **Qual é seu último conselho para os cientistas de dados em início de carreira?**

Conheça seus dados. Leva tempo – seis meses a um ano ou mais, se for um sistema complicado. Mas a qualidade dos seus dados é a base do seu universo. Se não conhecê-los, fará uma declaração muito bizarra sobre algo que seus dados simplesmente não atestam. Algumas pessoas dirão: “Tenho um número de cookies únicos que visitam meu site, e ele é igual ao número de pessoas únicas”. Mas não é verdade. E as pessoas que estão usando

vários dispositivos ou navegadores?

Para conhecer realmente seus dados, você precisa ser amigo das pessoas com conhecimento de domínio. Quando estava fazendo relatórios financeiros, fiz amigos no financeiro, pois assim eu aprendia as convenções que a contabilidade tem sobre como nomeiam coisas e a ordem de como as coisas são subtraídas. Talvez você tenha 50 milhões de páginas desse único IP, e outra pessoa perceberá que é a IBM. Você não saberá tudo, mas alguém provavelmente sim.

## **Resumo**

- Muitos tipos de empresas contratam cientistas de dados.
- Os empregos na ciência de dados variam, em grande parte, com base no setor, tamanho, história e cultura das equipes de cada empresa.
- É importante entender o tipo de empresa que você está considerando.

## CAPÍTULO 3

# Como aprender as competências

Este capítulo abrange:

- Diferentes formas de aprender ciência de dados
- Compreender o que é um bom curso acadêmico ou bootcamp
- Escolher o melhor caminho para você

Agora que você decidiu se tornar um cientista de dados, precisa aprender as competências! Não tema: pensar sobre como aprender as competências de um cientista de dados faz parte do como se tornar um cientista de dados. Para tanto, existem muitas formas, desde assistir a vídeos no YouTube até fazer faculdade. Muitas pessoas dirão que o caminho que escolheram é o único correto. Pior ainda, é fácil sentir-se sobrecarregado pela quantidade de coisas a aprender, como algoritmos, linguagens de programação e métodos estatísticos – e, depois, ainda se jogar em diferentes áreas. Só o fato de pensar sobre isso pode esgotar as energias!

A boa notícia é que existem apenas quatro métodos principais para aprender as competências necessárias. Cada método tem suas vantagens e desvantagens, mas, quando são comparados, fica em geral mais evidente qual abordagem é a melhor para você. No fim deste capítulo, você conseguirá entender os diferentes métodos e, após alguma reflexão, será capaz de decidir o melhor caminho para sua condição. Você consegue!

Os quatro métodos para aprender competências em ciência de dados neste capítulo são:

- Fazer um curso de pós-graduação em ciência de dados ou em um campo relacionado
- Participar de um bootcamp de ciência de dados (um curso de 8 a 15 semanas)
- Trabalhar com ciência de dados na sua atividade profissional

- Aprender por conta própria com cursos online e livros de ciência de dados

Vamos tratar de todos esses métodos neste capítulo.

**E se você não se formou na faculdade?**

Boa parte deste capítulo presume que você tenha se formado na faculdade, provavelmente em uma área técnica. Se não for o caso, não se preocupe; boa parte do capítulo ainda se aplica, mas você precisa fazer alguns ajustes ao lê-lo.

Se não se formou na faculdade, provavelmente seria melhor graduar-se primeiro antes de seguir as etapas do capítulo. Sua melhor aposta é realizar um curso em uma área técnica relacionada, que ensina algumas competências em ciências de dados, como matemática, estatística ou ciência da computação. Se escolher um desses cursos, tente fazer cadeiras eletivas para preencher as lacunas das competências em ciência de dados. Algumas universidades oferecem agora cursos de graduação em ciência de dados, o que faz de você um ótimo candidato para conseguir um emprego após a universidade. Após terminar seu curso, é provável que consiga um emprego em ciência de dados assim que se formar (especialmente se seguir a orientação das partes 1 e 2 deste livro). Você também pode optar por seguir as etapas adicionais listadas neste capítulo, tais como aprender por conta própria competências extras de ciência de dados ou trabalhar com ciência de dados em seu primeiro emprego.

Se você se formou em um curso que não é da área técnica, as orientações deste capítulo ainda servem. Contudo, pode ser uma ideia excelente iniciar um curso de pós-graduação em ciência de dados, pois um período de tempo mais longo de estudo dará mais tempo para correr atrás das suas competências técnicas. Talvez você se sinta inclinado a começar uma segunda graduação, agora em uma área técnica, mas evite isso completamente. Começar uma segunda graduação é muito caro e demorado, além do que você pode aprender de outras maneiras.

### **3.1 Curso universitário em ciência de dados**

Muitas universidades oferecem cursos de pós-graduação em ciência de dados, os quais tratam de uma mistura de tópicos de ciência da computação, estatística e negócios. Como são cursos de mestrado, via de regra levam dois anos e, nos EUA, custam 70 mil dólares ou mais. Além disso, como acontece com a maioria de cursos de pós-graduação, você pode escolher realizá-lo com mais calma se estiver trabalhando e/ou fazer as aulas na modalidade online. Embora muitas universidades ofereçam cursos de ciência de dados, dependendo de seus interesses você pode escolher iniciar um curso em ciência da computação, estatística, análise de negócios, pesquisa operacional ou em algo muito próximo à ciência de dados.

O bom de um curso de ciência de dados é sua abrangência; devido à duração do curso e à quantidade de tempo investido nele, você deve ter todo o conhecimento necessário para começar a carreira como cientista de dados júnior. Com cursos e projetos, terá experiência em estatística e métodos de machine learning, bem como programação na prática. Se ingressar no curso sem muita experiência em programação, será capaz de aprender ao longo do caminho (embora possa ter que fazer um ou mais cursos).

Cursos de pós-graduação em ciência de dados apresentam algumas desvantagens:

- São bem caros, tanto em termos de custo de mensalidade quanto de oportunidade, pois você não tem nem renda nem experiência de trabalho valiosa enquanto for um estudante em tempo integral.

Os cursos de pós-graduação são, em uma ordem de magnitude, mais



caros do que outras opções, em termos de dinheiro e tempo. Passar anos estudando antes de se sentir pronto para mudar de carreira é bastante tempo de vida, e, se decidir no meio do caminho que não quer ser um cientista de dados, não será possível recuperar o tempo e dinheiro despendidos.

- Se você vem de uma área relacionada à ciência de dados, como desenvolvimento de software, ou fez um curso de graduação significativo na área, uma pós-graduação pode abordar muitos temas que você já conhece. Isso significa que, em um curso longo, talvez aprenda poucas informações úteis e novas – uma enorme desvantagem que pode fazer com que o programa seja frustrante.
- Esses cursos são ministrados por professores acadêmicos, e a maioria deles passou uma carreira inteira na universidade, criando materiais de ensino que muitas vezes são diferentes do que se usam nas empresas. Um professor não atualizado pode usar linguagens antigas, como SPSS, por exemplo, ou não entender sobre ferramentas modernas, como controle de versão. Essa situação é sobremaneira comum em cursos fora da ciência de dados. Algumas universidades trazem pessoal de empresas para ministrar cursos, que podem não ter muita didática. É difícil saber se um curso usa técnicas modernas até começá-lo. Durante o processo seletivo, tente encontrar oportunidades para falar com alunos ou ex-alunos para ter uma ideia do curso e o quão útil ele é para uma carreira.

### **3.1.1 Escolha da universidade**

Ao começar a procurar cursos de pós-graduação em ciência de dados, você pode rapidamente ficar perdido diante da quantidade de opções oferecidas. Pior ainda: pode ficar repleto de propagandas de cursos diferentes e receber chamadas irritantes das instituições. Dependendo do quanto quiser se dedicar, é bom se candidatar a entre três e dez desses cursos. Se você se candidatar a poucos cursos, talvez não seja chamado para nenhum deles; caso se candidate a muitos, investirá tempo em demasia (e taxas).

Para decidir a que universidade se candidatar, considere o seguinte:

- *Se você vai ficar contente com a localização e o estilo de vida [muito*

*importante]* – provavelmente, você pesquisará cursos de pós-graduação em vários lugares, mas sua vida fora da universidade será bastante diferente dependendo da cidade. Se o clima, proximidade a amigos ou custo de vida não forem bons, não importa o quão bom seja o curso, porque ficará infeliz com ele.

- *Quais tópicos o curso aborda [importante]* – como a ciência de dados é muito nova, as universidades podem ter uma grade curricular bastante diferente. Essa situação é especialmente complicada dependendo do departamento em que o curso se encontra. Um curso de ciência de dados baseado em ciência da computação terá como foco métodos e algoritmos, por exemplo; já um curso em uma escola de administração terá foco em aplicações e será baseado em estudos de caso. Verifique se o material do curso aborda seus pontos fracos em relação a competências (consulte o Capítulo 1).
- *Quantos projetos o curso oferece [importante]* – quanto mais projetos um curso oferecer, mais você aprenderá sobre como a ciência de dados funciona na prática e mais bem preparado ficará para trabalhar em uma empresa. (Os projetos são tratados em maior profundidade no Capítulo 4.) Projetos significativos também são ótimos para colocar no currículo, o que pode ajudá-lo a conseguir um estágio durante o curso ou mesmo um primeiro emprego mais tarde.
- *Para aonde vão os formados [importante]* – muitas vezes, uma universidade tem estatísticas sobre aonde os alunos vão depois da graduação, como a porcentagem que vai seguir a carreira acadêmica ou trabalhar para as empresas Fortune 500. Essas estatísticas podem ser informativas, mas as universidades mostram os números que são melhores para elas, mesmo que enganosos (o que é irônico, pois entender do que se constitui uma estatística enganosa é uma competência que aprenderá como cientista de dados). Se possível, tente falar com alguns alunos antigos do curso pelo LinkedIn para ter uma perspectiva imparcial sobre como os graduados se saíram. Especialmente se quiser trabalhar em uma grande corporação, dá para pesquisar quais empresas recrutam diretamente nessa universidade. Ainda é possível se candidatar a um emprego se essas empresas não

recrutarem, mas pode receber uma menor consideração.

- *Bolsa [rara, mas muito importante]* – em raras circunstâncias as universidades oferecem bolsas para alunos de mestrado, cobrindo o curso e às vezes pagando um valor se você for assistente de ensino em uma disciplina. Se oferecerem uma bolsa, recomendamos muito que a aceite; não ter de pagar o curso e ainda receber é financeiramente uma opção muito melhor do que ter de pagar as despesas por conta própria. Se envolver dar aulas, você também terá a oportunidade de ser forçado a aprender a se comunicar para várias pessoas, o que será útil na sua carreira de ciência de dados. A desvantagem é que dar aulas demanda muito tempo, o que irá distraí-lo de seus estudos.
- *A proximidade do curso com as empresas da área [médio-importante]* – se a universidade faz muitos trabalhos com as empresas locais, especialmente as de tecnologia, a universidade está conectada com a comunidade. Essa conexão tornará mais fácil conseguir um estágio ou um emprego, além de oferecer materiais mais interessantes para as aulas. Também reduz a chance de ter professores que estão sem contato com os métodos utilizados fora da universidade.
- *Requisitos de admissão [não muito importante]* – algumas universidades exigem que você tenha feito certas disciplinas para ser aprovado. A maioria dos cursos exige que você tenha cursado disciplinas de matemática, como álgebra linear, além de programação, como Introdução a Java. Se faltarem algumas disciplinas necessárias, talvez você consiga liberação do requisito ou fazer disciplinas de compensação quando estiver no curso. Se você não tiver nenhum pré-requisito ou se a universidade exigir uma graduação específica (como ciência da computação), talvez o curso não seja adequado para você.
- *Prestígio da universidade [não é tão importante]* – a menos que você seja aceito em uma universidade de grande prestígio, os empregadores não se importarão se a universidade é tão conhecida. O prestígio é importante principalmente se estiver buscando seguir a carreira acadêmica em vez de trabalhar em empresas, mas, nesse caso, você deve pensar em um doutorado e não só em um mestrado (e também ler um

livro diferente deste). O prestígio só é útil para redes de contato de egressos que essas universidades proporcionam.

- *Seu orientador [muito importante, mas...]* – se o curso de pós-graduação que você está considerando tiver uma dissertação ou tese, você terá um orientador na universidade. Ter um orientador com um estilo de trabalho e área de interesse similares aos seus, além de não ser uma pessoa abusiva, facilmente fará a diferença entre ser aprovado ou não no curso. Infelizmente, antes de ingressar na universidade, é muito difícil saber qual orientador será melhor, e mais ainda conhecer a personalidade dele. Assim, embora esse ponto seja extremamente importante, não há muito o que fazer para tomar uma decisão com base nele. Se, no entanto, o curso for inteiramente baseado em trabalhos ou tiver apenas um projeto final, um orientador não terá tanta importância.

Quando estiver pensando nas universidades, tente fazer uma planilha que liste como as universidades se saem em cada um desses pontos. Quando você tem todos os dados, pode ser difícil classificar as escolas de maneira objetiva. Como saber de fato se uma universidade em uma cidade onde você odiaria viver, mas que está bem conectada às empresas, é melhor ou pior do que uma universidade em uma cidade que você adoraria viver, mas que não tem um projeto? Recomendamos que se esqueça da ideia de encontrar a “melhor” de maneira objetiva. Em vez disso, agrupe as universidades em “amo”, “gosto” e “ok” e, depois, candidate-se apenas àquelas que você ama ou gosta.

## **Cursos de pós-graduação online**

Cada vez mais universidades estão oferecendo cursos de pós-graduação online, tornando possível aprender tudo o que você precisa saber sem ter que ir a um campus universitário. O benefício óbvio desses cursos é que instituir cursos online é mais conveniente do que ter que passar horas toda semana indo à universidade. Além disso, cursos online não têm mais o estigma que tinham no início, então não é preciso se preocupar se os empregadores considerarão legítima ou não sua formação. A desvantagem está na dificuldade de envolvimento com o curso e com o material didático caso faça tudo online. Será mais difícil interagir com seus professores quando tiver dúvidas, mas será mais fácil prestar menos atenção e não fazer as tarefas. De certa forma, a conveniência de um curso online também pode ser um tiro no pé: não há tanto incentivo para se dedicar a ele. Se pensa que tem a capacidade de permanecer comprometido e focado em um curso online, pode ser uma ótima escolha; apenas cuidado com os riscos.

### **3.1.2 Como ingressar em um curso universitário**

Para ingressar em um curso universitário, é preciso candidatar-se. Caso esteja familiarizado com o processo seletivo de acesso a cursos de pós-graduação, o procedimento para um mestrado em ciência de dados será similar aos demais. A primeira etapa é reunir os documentos para inscrição. Normalmente, as universidades anunciam no início do ano como se candidatar, incluindo prazos e materiais necessários. Nos EUA, os processos seletivos de cursos de pós-graduação normalmente requerem o seguinte:

- *Uma carta de intenções de uma a duas páginas* descrevendo por que você é um bom candidato ao curso. Nessa carta, concentre-se o máximo possível na razão pela qual você seria uma boa contribuição para o curso. Coisas como ter experiência em algumas competências exigidas pela ciência de dados ou exemplos de trabalho relacionado que você desenvolveu são extremamente úteis. Tente evitar clichês como “tenho interesse em ciência de dados desde criança”. Há muitos recursos para escrever um bom texto, e a universidade onde concluiu a graduação talvez tenha um departamento para ajudá-lo nessa tarefa.

- *Histórico escolar* da sua graduação para comprovar que você tem os pré-requisitos necessários para o curso. O site da sua universidade deve ter instruções sobre como obter seu histórico escolar, mas lembre-se de que geralmente custa dinheiro e leva uma semana ou mais para ser entregue. Não deixe essa tarefa para o último minuto!
- *Pontuação do Graduate Record Examination (GRE, Exame de Registro do Graduado)* alcançando um nível mínimo de competências verbais e matemáticas. Em teoria, o GRE de matemática deve ser fácil para qualquer pessoa que ingresse em um programa de ciência de dados, pois a matemática é a base da ciência de dados. No entanto, muitas pessoas não viram questões complicadas de matemáticas desde o ensino médio, então é uma boa ideia estudar. A prova verbal pode ser mais difícil e exigir mais preparo. Os GREs exigem que você vá a um local especial e pode ser complicado se programar. Portanto, seja proativo e tente fazer o exame logo. Se o inglês não for seu idioma nativo, provavelmente precisará obter uma pontuação mínima no Test of English as a Foreign Language (TOEFL, teste de inglês como língua estrangeira) ou no International English Language Testing (IELTS, sistema internacional de testes de língua inglesa).
- *Três cartas de recomendação* afirmando por que você seria bom para esse curso de pós-graduação. Essas cartas podem ser de professores que você teve ou de pessoas como seu chefe caso seu trabalho esteja relacionado à ciência de dados. Preferencialmente, quem escrever a carta deve ser capaz de falar sobre por que você seria um bom cientista de dados; portanto, essas pessoas precisam ter visto você se saindo bem. Tente evitar professores universitários que digam apenas “esta pessoa tirou um A na minha disciplina” e empregadores que não podem dizer muito sobre seu trabalho em um ambiente técnico. Se você é um estudante de graduação que está lendo este livro, agora pode ser um bom momento para conhecer melhor seus professores, conversando com eles, e participando de seminários e clubes acadêmicos.

Esses materiais demandam tempo para serem reunidos, e, se estiver se candidatando a muitas universidades ao mesmo tempo, reunir todo esse material pode acabar virando uma atividade de tempo integral. Nos EUA, a

maioria das inscrições está prevista para dezembro e fevereiro, e você teria um retorno por volta de fevereiro ou março. Se for aceito, você tem até abril para decidir se quer ingressar no curso. Quando receber suas cartas de aceite, não se preocupe muito com a universidade que é a “melhor” – basta escolher uma onde acha que ficará feliz!

### **3.1.3 Resumo de cursos acadêmicos**

Em suma, cursos de pós-graduação em ciência de dados são uma boa opção para pessoas que querem uma instrução abrangente e que tenham recursos para tanto. Essas pessoas podem vir de uma área em que não programaram muito nem fizeram bastante trabalho técnico, como marketing. Um curso de pós-graduação permite que aprendam todos os componentes da ciência de dados em um ritmo aceitável de transição.

Os cursos de pós-graduação não são bons para as pessoas que já têm muitas das competências necessárias; os cursos são muito longos e caros demais para valerem a pena, ademais, os professores não são especialistas do setor, portanto, o pouco novo conhecimento que transmitem pode nem mesmo ser muito relevante. Pode ser preciso começar a estagiar em empresas durante o curso de pós-graduação para melhorar o curso.

Se estiver pensando que precisa de treinamento extensivo antes de ser um cientista de dados, então siga em frente. Comece buscando universidades que tenham o curso de pós-graduação de que você gosta. Se achar que estudar um curso desses vai demandar muito trabalho e que deve haver uma forma mais fácil, considere as opções das próximas seções.

**Preciso de um doutorado para conseguir um emprego em ciência de dados?**

Provavelmente não.

Doutorados são formações que demandam muitos anos para concluir e focam em estudantes que estão se preparando para a docência. O aluno tem que passar anos fazendo pesquisas para encontrar um novo método ou solução para um problema e que seja ligeiramente melhor do que um anterior. Você publica em revistas científicas e avança em pesquisas de última geração em uma área extremamente específica. Mas como os capítulos 1 e 2 demonstraram, pouco do trabalho que um cientista de dados faz é como uma investigação acadêmica. Um cientista de dados preocupa-se muito menos em encontrar uma solução elegante e de ponta e muito mais em encontrar rapidamente algo que seja suficientemente bom.

Um bom número de vagas de emprego na ciência de dados requer um doutorado. Mas as competências adquiridas em um doutorado raramente são necessárias para o trabalho; normalmente, o requisito de doutorado é um sinal da empresa de que a posição é considerada de prestígio. O conteúdo que você pode aprender em um mestrado ou uma graduação irá adequá-lo para a grande maioria dos empregos em ciência de dados.

Além disso, começar um doutorado tem um custo enorme de oportunidade. Se levar sete anos para se formar, você poderia ter trabalhado em uma empresa durante esse período, ficando melhor em ciência de dados e ganhando muito mais dinheiro.



Você pode fazer um doutorado e, depois, tornar-se um cientista de dados, mas não deixe ninguém dizer que você precisa dessa formação.

## **3.2 Participação de bootcamps**

Um *bootcamp* é um curso intensivo de 8 a 15 semanas, oferecido por empresas como a Metis e a Galvanize. Durante o bootcamp, você passa mais de oito horas diárias aprendendo competências da ciência de dados, ouvindo palestrantes do setor e trabalhando em projetos. No fim do curso, você normalmente apresentará um projeto final a uma sala cheia de pessoas de empresas querendo contratar cientistas de dados. Em teoria, sua apresentação fará com que consiga uma entrevista e, então, um emprego.

Bootcamps ensinam uma quantidade significativa de coisas em um tempo muito curto, o que significa que podem ser ótimos para pessoas que têm a maioria das competências necessárias para a ciência de dados, mas necessitam de um pouco mais. Considere alguém que trabalha como neurocientista e fez algo de programação como parte de seu trabalho. Um bootcamp de ciência de dados poderia ensinar-lhe tópicos como regressões logísticas e bancos de dados SQL. Com a base científica mais esses princípios básicos, a pessoa estaria pronta para conseguir um emprego em ciência de dados. Às vezes, a melhor parte de um bootcamp não é o conhecimento em si, mas a confiança que você adquire de que pode de fato trabalhar na área.

### **3.2.1 O que você aprende**

Um bom bootcamp tem um programa altamente otimizado para ensinar exatamente o que você precisa saber para conseguir um trabalho em ciência de dados – e nada mais. O programa vai além das competências técnicas, incluindo oportunidades de trabalhar em projetos e construir redes de contato. As seções a seguir oferecem mais detalhes sobre o que você deve esperar que um bootcamp aborde.

### **Competências**

Bootcamps são grandes complementos à formação existente. Ao fazer um

bootcamp, é possível conseguir rapidamente um emprego em ciência de dados, sem passar dois anos em um curso (se fosse fazer um mestrado). Esse fato pode ser especialmente atraente se já tiver um mestrado em uma área de ciências que não seja a de dados. As competências que você normalmente aprende em um bootcamp são:

- *Estatística básica* – essa competência inclui métodos para fazer previsões com dados, como regressões lineares e logísticas, bem como métodos de teste que você poderia usar no trabalho, como testes t. Devido à duração muito limitada, as razões pelas quais esses métodos funcionam não serão muito aprofundadas, mas você aprenderá muito sobre como utilizá-los.
- *Métodos de machine learning* – o programa abordará algoritmos de machine learning, como florestas aleatórias (random forests) e máquinas de vetores de suporte (support vector machines), além de mostrar como usá-los dividindo dados em grupos de treinamento e de teste e empregando a validação cruzada (cross validation). Você pode aprender algoritmos para casos específicos, como o processamento de linguagem natural ou mecanismos de busca. Se nenhuma dessas palavras faz sentido para você, um bootcamp pode ser uma boa opção!
- *Programação intermediária em linguagem R ou Python* – você aprenderá conceitos básicos de como dados são armazenados em quadros de dados e como manipulá-los, resumindo, filtrando e plotando dados. Também aprenderá a usar os métodos estatísticos e de machine learning dentro do programa escolhido. Embora possa aprender R ou Python, provavelmente não aprenderá os dois, então terá de aprender a outra linguagem depois de terminar o bootcamp se precisar dela no seu primeiro emprego.
- *Casos de uso no mundo real* – você aprenderá não apenas os algoritmos, mas também onde as pessoas usam esses algoritmos, com uma regressão logística para prever quando um cliente deixará de assinar um produto, por exemplo, ou com um algoritmo de cluster para segmentar os clientes para uma campanha de marketing. Esse conhecimento é extremamente útil para conseguir um emprego, e as perguntas sobre casos de uso

aparecem com frequência nas entrevistas.

## **Projetos**

Bootcamps têm um currículo altamente baseado em projetos. Em vez de ouvir palestras oito horas por dia, você passará a maior parte do tempo trabalhando em projetos que podem ajudá-lo a entender a ciência de dados e a começar a montar seu portfólio na área (assunto do Capítulo 4). É uma enorme vantagem em relação ao mundo acadêmico, pois suas competências estarão alinhadas com o que você necessita para ter sucesso no setor, o que é muitas vezes similar ao trabalho baseado em projetos.

Em um projeto, você primeiramente coleta dados. É possível coletar esses dados usando uma API web criada por uma empresa para extrair seus dados, vasculhando sites para obter informações deles ou usando conjuntos de dados públicos de lugares como sites do governo. Depois, carregará os dados em R ou Python, escreverá scripts para manipular os dados e executará modelos de machine learning nos mesmos. Em seguida, usará os resultados para criar uma apresentação ou relatório.

Nenhuma dessas etapas no projeto requer um bootcamp. De fato, o Capítulo 4 deste livro é inteiramente sobre como fazer projetos de ciência de dados por conta própria. Dito isso, ter um projeto como parte de um bootcamp significa que haverá instrutores o orientando e o auxiliando se as coisas correrem mal. É difícil manter-se motivado se estiver trabalhando sozinho e é fácil ficar desestimulado se não houver uma pessoa para quem pedir ajuda.

## **Uma rede de contatos**

Muitas pessoas saem de bootcamps e começam uma carreira bem-sucedida em lugares como Google e Facebook. Os bootcamps mantêm redes de contato de alunos antigos que podem ser usadas para abrir a porta dessas empresas. O bootcamp pode trazer palestrantes da ciência de dados durante o curso, bem como pessoas de empresas para assistir às suas apresentações finais. Todas essas pessoas podem servir como conexões para ajudá-lo a conseguir um emprego nas suas empresas. Ter portas abertas em empresas com vagas de ciência de dados pode fazer toda a diferença quando se trata

de encontrar um emprego, por isso vale a pena salientar esse aspecto positivo de bootcamps.

Além de conhecer pessoas durante o curso, você pode usar ferramentas como o LinkedIn para entrar em contato com alunos do bootcamp. Essas pessoas podem ser capazes de ajudá-lo a encontrar um emprego nas empresas delas ou, pelo menos, indicar uma empresa que seria uma boa opção para você.

Para todas essas conexões, você precisa ser proativo, o que pode significar falar com palestrantes ao final das apresentações e enviar mensagens em redes sociais para pessoas com as quais você nunca falou antes. Esse processo pode ser assustador, especialmente se não se sentir confortável interagindo com pessoas que não conhece, mas é necessário tirar o máximo de proveito do bootcamp. Consulte o Capítulo 6 para conhecer sugestões sobre como escrever uma solicitação para rede de contato de forma eficaz.

### **3.2.2 Custo**

Uma desvantagem significativa de um bootcamp em comparação ao autoaprendizado é o custo: as taxas geralmente são de cerca de 15 a 20 mil dólares. Embora você possa conseguir bolsas de estudo para cobrir parte da mensalidade, também tem de considerar o custo da oportunidade de não ser capaz de trabalhar em tempo integral (e provavelmente nem mesmo meio período) durante o programa. Além disso, provavelmente ficará buscando emprego por vários meses após o bootcamp. Você não conseguirá se candidatar durante o bootcamp porque estará muito ocupado e ainda não terá aprendido as competências, e inclusive um processo bem-sucedido de candidatura a um emprego em ciência de dados pode levar meses até a data de início. No total, você pode acabar ficando desempregado por seis a nove meses por causa do bootcamp. Se puder aprender ciência de dados por conta própria durante o seu tempo livre ou no trabalho, será possível continuar trabalhando e não pagar o curso, podendo economizar muito dinheiro.

### **3.2.3 Escolha do curso**

Dependendo de onde você reside, talvez tenha apenas algumas opções para

bootcamps. Se quiser fazer um bootcamp presencial, mesmo uma cidade grande pode oferecer apenas poucos programas. Se não mora em uma cidade grande e quer fazer um bootcamp, poderá ter que se mudar temporariamente para um centro maior, o que aumentaria o custo do programa e dificultaria sua vida.

Outra opção são os bootcamps online. No entanto, tenha cuidado: como acontece com os cursos de pós-graduação, um dos benefícios dos bootcamps presenciais é ter pessoas à sua volta para o motivá-lo e manter o foco. Se fizer um curso online, você perde essa vantagem, o que pode tornar um bootcamp na versão online de 20 mil dólares um curso igual que dá para encontrar de graça ou mais barato.

Ao escolher entre os bootcamps em sua área, considere conferir as salas de aula, falar com alguns dos instrutores e ver onde você se sentirá mais confortável. Mas cuidado: tanto na universidade como nos bootcamps, muita gente está buscando ganhar dinheiro fácil daqueles que desejam se tornar cientistas de dados. Se não tiver cuidado, poderá terminar em um programa que não o ajudará a conseguir um emprego e que o deixará com dívidas. Para os bootcamps, é extremamente importante falar com alunos antigos. Você encontra graduados bem-sucedidos no LinkedIn? Em caso afirmativo, converse com essas pessoas para saber como se sentem com relação à experiência que tiveram. Se não conseguir encontrar pessoas no LinkedIn que passaram pelo programa, esse fato é um grande alerta.

### **3.2.4 Resumo dos bootcamps de ciência de dados**

Os bootcamps podem ser ótimos programas para pessoas que querem mudar de carreira e já conhecem alguns dos princípios da ciência de dados. Também podem ser úteis para aquelas que estão recém saindo da universidade e querem ter alguns projetos de ciência de dados no portfólio enquanto procuram emprego. No entanto, os bootcamps não são pensados para te levar do zero ao cem. A maioria deles tem exigências de admissões competitivas, sendo necessário ter uma base de noções de programação e estatística para o ingresso e, assim, aproveitar o programa ao máximo.

### 3.3 Trabalhar com ciência de dados dentro de sua empresa

Você pode estar em um trabalho relacionado à ciência de dados. Um método incomum, mas muitas vezes bastante eficaz de aprender ciência de dados é começar a atuar cada vez mais com ciência de dados no trabalho. Talvez você trabalhe na área administrativa, trazendo essa perspectiva aos relatórios de ciência de dados, podendo começar a fazer seus próprios gráficos. Ou talvez você trabalhe no financeiro, elaborando planilhas que poderiam ser feitas em linguagem R ou Python.

Considere o caso hipotético de Luísa, que trabalha há muitos anos no departamento de pesquisa de mercado, fazendo pesquisas sobre clientes e usando uma interface gráfica de usuário (GUI) de pesquisa de mercado para agregar os resultados de pesquisa. Sua formação é em sociologia e programou um pouco nos anos de graduação. Com frequência trabalha com o departamento de ciência de dados, passando adiante dados de pesquisa e ajudando os cientistas de dados a entendê-los para que possam usá-los em modelos. Com o tempo, Luísa começa a fazer pequenos trabalhos para a equipe de ciência de dados – um pouco de extração de recursos em R aqui, outro pouco de criação de visualizações ali. Em pouco tempo, a equipe de ciência de dados está cada vez mais dependente de Luísa. Nesse meio tempo, as competências de programação e ciência de dados da equipe estão melhorando. Depois de um ano, ela junta-se à equipe de ciência de dados em tempo integral, deixando a pesquisa de mercado para trás.

Tentar fazer algo de ciência de dados no trabalho é um método excelente, pois é de baixo risco e tem motivação agregada. Você não está tentando fazer um bootcamp caro ou um curso de pós-graduação no qual precisará deixar seu trabalho. É só uma questão de trabalhar um pouco mais, com ciência de dados, quando conseguir. E o fato de estar fazendo ciência de dados no seu próprio trabalho é motivador porque o trabalho que realizará é valioso para os demais. Com o tempo, será possível realizar mais trabalho de ciência de dados até fazer apenas isso, ao contrário de fazer um curso e, então, de repente, trocar de emprego.

A ex-pesquisadora de mercado e agora cientista de dados, Luísa, tinha

boas vantagens:

- Relações existentes com o departamento de ciência de dados que podia orientá-la.
- Entendimento de conceitos básicos de programação e visualização de dados.
- Motivação suficiente para aprender técnicas de ciência de dados no trabalho.
- O departamento de ciência de dados foi capaz de passar pequenos projetos e, com o tempo, esses projetos cresceram, permitindo que Luísa se tornasse uma cientista de dados.

Quando você está tentando fazer mais trabalho de ciência de dados na empresa, procure locais onde possa encontrar pequenos projetos de ciência de dados e pessoas para ajudá-lo. Algo tão simples quanto criar ou automatizar um relatório pode ensinar muito sobre ciência de dados.

Uma observação importante caso siga nessa direção: nunca se torne um fardo para alguém. Os inconvenientes podem ser muito óbvios, como pedir várias vezes às pessoas que criem conjuntos de dados limpos para você, ou, menos óbvios, como solicitar frequentemente a alguém que revise o trabalho que fez. Também é possível criar um fardo acidental adicionando novas ferramentas à equipe. Se estiver no financeiro e todos usarem o Microsoft Excel, exceto você (agora usando a linguagem R), acabou de tornar a gestão da sua equipe mais complicada. Mesmo a ação de pedir trabalho a alguém constitui-se em um fardo, pois ele precisa encontrar algo para você fazer. Tenha cuidado ao aprender essas competências, a fim de não criar problemas para seus colegas.

## **Duas perspectivas**

*O que você diz:* “Estou feliz em ajudar de alguma maneira, avise-me como! Obrigado(a)!”

*O que você acha que entendem:* “Sou uma pessoa que está animada para trabalhar com você. Pode me passar esse projeto bacana, mas simples, que está na gaveta há tanto tempo, e eu o farei para você!”



*O que realmente entendem:* “Olá! Quero ser útil, mas não tenho ideia alguma do que você precisa. Também não sei quais são minhas competências em relação ao seu trabalho; por isso, boa sorte em encontrar uma tarefa para mim. Além disso, se encontrar uma tarefa perfeita para mim, provavelmente terá que revisá-la várias vezes antes de ficar boa. Todo esse trabalho ocupará suas já poucas horas disponíveis. Obrigado(a)!”

Para esse caminho funcionar, você precisa empregar algumas estratégias importantes:

- *Ser proativo* – quanto mais fizer o trabalho antes que as pessoas peçam, mais independente se tornará sem ser um fardo. A equipe de ciência de dados pode ter uma tarefa, como rotular dados ou fazer um relatório simples, o que requer muito tempo e não é muito interessante. Você pode oferecer ajuda para esse trabalho. Tenha cuidado em mergulhar na tarefa e fazê-la sozinho: você pode acabar fazendo a tarefa de uma forma que não agrega valor e que a equipe terá que refazê-la. No entanto, se começar a tarefa e perguntar o que outras pessoas acham, é possível economizar muito tempo da equipe.
- *Escolha uma competência nova de cada vez* – não tente aprender tudo sobre ciência de dados de uma vez só. Encontre uma competência que queira aprender com o trabalho e, então, a aprenda. Você pode querer aprender como fazer relatórios com a linguagem R, por exemplo, pois a equipe de ciência de dados faz isso o tempo todo. Ao encontrar um projeto pequeno para ajudar a equipe, você pode aprender essa competência e adicioná-la à sua caixa de ferramentas. A partir daí, é possível aprender uma competência diferente em ciência de dados.
- *Deixe claras suas intenções* – ficará bastante óbvio que você está tentando pegar trabalho extra para aprender a ser um cientista de dados. Se for proativo e deixar a equipe de ciência de dados saber de seu interesse em aprender mais, a equipe poderá planejar tê-lo como ajuda. Além disso, a equipe será mais compreensiva sobre sua inexperiência, porque um dia também eles eram novatos e estavam aprendendo.
- *Evite ser intrometido* – ajudar uma pessoa a se tornar um cientista de

dados dá bastante trabalho, e muitas vezes as equipes de ciência de dados já estão sobrecarregadas. Se acha que a equipe não tem tempo ou capacidade para ajudá-lo, não leve para o lado pessoal. Embora seja possível aparecer de vez em quando se achar que a equipe não está entrando em contato, se for persistente demais com seus pedidos, todos podem se sentir desconfortáveis. A equipe verá você não como um recurso potencial, mas mais como um incômodo.

### **Quando não há oportunidades**

Você pode se encontrar em uma situação em que não existem oportunidades para utilizar a ciência de dados no seu cargo atual. É possível que as restrições de seu trabalho o impeçam de usar a linguagem R ou Python ou de tentar implementar técnicas de ciência de dados. Nessas situações, poderá ter de tomar medidas drásticas. Abandonar seu emprego para fazer um bootcamp ou um curso de pós-graduação é arriscado, mas eficaz para sair do cargo atual e ir para a ciência de dados. Você poderia tentar também aprender por conta própria em seu tempo livre, mas esse método tem toneladas de desvantagens (ver a Seção 3.4). Outra opção é tentar encontrar um novo emprego em seu campo que lhe permita aprender mais nesse cargo. Mas não há garantia alguma de que, quando chegar no trabalho novo, terá a flexibilidade prometida.

Nenhuma dessas opções é fácil, mas, infelizmente, é a realidade da situação. Demanda esforço conseguir um cargo em ciência de dados, mas pode valer a pena.

### **3.3.1 Resumo de aprender no trabalho**

Aprender no trabalho pode ser uma maneira eficaz de se tornar um cientista de dados, desde que você tenha um emprego em que pode aplicar competências de ciência de dados e que haja pessoas que possam orientá-lo. Se tudo se alinhar, é um ótimo caminho, mas, para muitos, não é bem assim. Se acredita que esse caminho é viável para você, recomendamos segui-lo. Muitas vezes, certos cargos não permitem aprender no trabalho, por isso, aproveite a oportunidade se a tiver.

## **3.4 Aprender por conta própria**

Um número enorme de livros aborda a ciência de dados (como este aqui), além de muitos cursos online. Esses livros e sites prometem ensinar os princípios básicos da ciência de dados, bem como as competências técnicas em profundidade através de um meio (e a um preço) que é prático. Esses cursos e livros – bem como todos os blogs, tutoriais e respostas do Stack Overflow sobre ciência de dados – podem fornecer uma base suficiente para que as pessoas possam aprender por conta própria.

Esses materiais de aprendizagem autodidata são ótimos para aprender competências individuais. Se quiser compreender como fazer a deep learning (aprendizado profundo), por exemplo, um livro pode ser uma ótima maneira. Ou se quiser aprender conceitos básicos em linguagem R e Python, pode fazer um curso online para começar.

Aprender ciência de dados inteiramente por conta própria em cursos e livros online e de forma autodidatas é como aprender a tocar um instrumento com vídeos do YouTube ou aprender qualquer outra coisa sem um professor: o valor desse método depende da sua perseverança. Aprender novas competências pode levar centenas ou milhares de horas. É muito difícil colocar milhares de horas na ciência de dados quando as melhores compilações de TikTok estão logo ali na outra janela. Também é difícil saber por onde começar. Se quiser aprender tudo da ciência de dados, quem dirá qual livro você deve ler primeiro (como este aqui)?

Aprender por conta própria significa que você não tem um professor ou um modelo a seguir. Por não ter um professor para quem fazer perguntas, como faria em um curso acadêmico ou em um bootcamp, você não saberá quando está fazendo algo errado ou o que fazer a seguir. O tempo gasto sem ter uma direção ou seguir um caminho incorreto constitui-se em um obstáculo para entender o assunto. A melhor maneira de compensar a falta de um professor é encontrar uma comunidade de pessoas para fazer perguntas. Um ótimo exemplo é o programa TidyTuesday (<https://github.com/rfordatascience/tidyesday>), iniciado por Thomas Mock; todas as terças-feiras, os aspirantes e praticantes de ciência de dados usam linguagem R para lidar com um problema da ciência de dados.

Se decidir seguir o caminho autodidata, é importante ter algum trabalho construtivo a fazer. Ler livros e assistir a vídeos é ótimo, mas você aprende muito mais fazendo seu próprio trabalho de ciência de dados e aprendendo com esse trabalho. Em outras palavras, ler livros sobre bicicletas pode ser educativo, mas nunca aprenderá a andar de bicicleta sem andar de bicicleta. Certifique-se de encontrar um projeto que deseja fazer, como analisar um conjunto de dados e encontrar resultados interessantes sobre ele, criar um modelo de machine learning e API ou usar uma rede neural para gerar texto. No Capítulo 4, vamos entrar em muito mais detalhes sobre esse tipo

de projeto. Para outros métodos de aprendizado em ciência de dados, os projetos podem ser apenas para montar portfólio, mas, para o aprendizado autodidata, os projetos são cruciais para a aprendizagem.

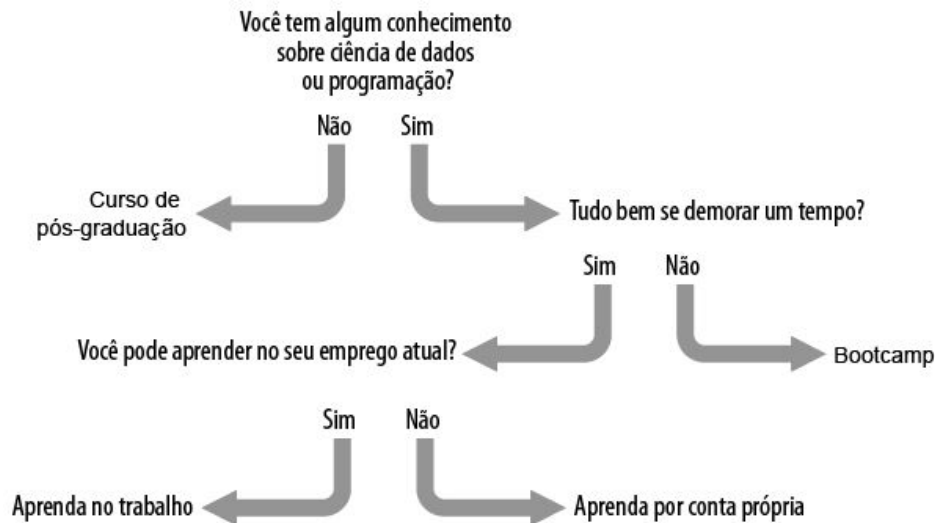
### **3.4.1 Resumo de aprender por conta própria**

Aprender sozinho é difícil – possível, mas difícil. Você precisa ser capaz de descobrir a ordem em que quer aprender as coisas, manter-se motivado o suficiente para aprender as competências e fazer tudo sem ter um mentor ou professor para ajudá-lo. Também terá mais dificuldade de exibir suas qualificações em um currículo do que se usar outros métodos. Tal método é nossa forma menos recomendada de se tornar cientista de dados devido ao número de coisas que podem correr mal e ao número de pessoas que não conseguem manter-se focadas. Se quiser pegar uma única competência ou tecnologia, seguir esse caminho pode ser mais viável, mas aprender tudo que precisa para ser um cientista de dados é uma trajetória difícil.

## **3.5 Escolha**

Como escolher entre esses quatro caminhos muito diferentes para a ciência de dados? O processo é diferente para todos, mas propomos responder a três perguntas (Figura 3.1):

1. *Você já tem algum conhecimento de ciência de dados?*  
Especificamente, você já programou em pelo menos uma linguagem de programação além de algum curso básico? Sabe como fazer consultas em um banco de dados SQL? Sabe, por exemplo, o que são regressões lineares?



*Figura 3.1 – Fluxo de decisão sobre qual caminho escolher para a educação em ciências de dados.*

- a. Se sua resposta for “não, tenho muito a aprender”, é provável que um curso acadêmico, como um de pós-graduação, seja mais adequado para você. Poderá aprender todos esses tópicos no curso, em um período de tempo suficientemente longo para ficarem fixados.
  - b. Se sua resposta for “sim, sei dessas coisas”, passe para a pergunta 2.
2. *Você sente-se confortável em tirar um ano ou mais para aprender as competências necessárias em ciência de dados em vez de ter os custos de ficar desempregado por seis a nove meses para começar a trabalhar com ciência de dados mais rapidamente?* É difícil aprender novas competências com rapidez quando se concentra apenas na aprendizagem; é ainda mais difícil fazer isso trabalhando em tempo integral. Considera tudo bem se o caminho levar mais tempo para que possa continuar trabalhando em tempo integral?
- a. Se sua resposta for “não, tem que ser rápido”, faça um bootcamp. Em três meses você saberá toneladas sobre ciência de dados e estará pronto para embarcar em sua busca por um novo emprego, que pode levar de três a seis meses a mais.
  - b. Se sua resposta for “sim, quero ocupar meu tempo”, passe para a pergunta 3.

3. *Você pode aprender ciência de dados no seu trabalho?* Você pode fazer coisas de ciência de dados no seu cargo atual, como, por exemplo, uma análise, armazenar alguns dados em SQL ou tentar usar as linguagens R ou Python? Há uma equipe que possa orientá-lo ou lhe repassar pequenas tarefas?

- a. Se sua resposta for “sim, posso aprender no trabalho”, tente isso e use seu emprego como trampolim para a ciência de dados.
- b. Se sua resposta for “não, meu trabalho não oferece oportunidades”, é hora de mergulhar nos livros e nos cursos online.

Embora essas perguntas sejam um ponto de partida, não é preciso tomar uma decisão final. Você pode começar a ler livros de forma independente, e, se achar que quer seguir com mais rapidez, troque para um bootcamp. Você também pode fazer um curso de pós-graduação à noite enquanto tenta fazer ciência de dados no trabalho. Não há uma resposta perfeita; o que importa é encontrar uma abordagem que funcione para você. Se algo não estiver funcionando, mude até que funcione.

Depois de escolhida a sua rota, está na hora de segui-la! Inscreva-se no curso de pós-graduação, faça aquele bootcamp ou compre uns livros e comece a ler. Para os fins deste livro, presumimos que o tempo passou e você conseguiu aprender as competências fundamentais necessárias para ser um cientista de dados. Nos próximos capítulos, usará essas competências para criar um portfólio de ciência de dados que possa ajudá-lo a conseguir seu primeiro trabalho na área.

## **3.6 Entrevista com Julia Silge, cientista de dados e engenheira de software da RStudio**

Julia Silge é conhecida pelos seus posts de blog sobre ciência de dados, juntamente com o pacote `tidytext` que ela e David Robinson desenvolveram, que é um marco de PLN em linguagem R e foi baixado mais de 700 mil vezes. Ela e Robinson são autores do livro *Text Mining with R: A Tidy Approach* (O’Reilly, Mineração de Texto com R: uma abordagem organizada, em tradução livre). Julia trabalhou durante vários anos como cientista de dados na Stack Overflow e agora desenvolve ferramentas de

machine learning de código aberto na RStudio.

## **Antes de se tornar cientista de dados, você trabalhou na universidade. Como as competências aprendidas lá a ajudaram como cientista de dados?**

Quando estava fazendo pesquisa como acadêmica, alguns dias eram passados na coleta de dados do mundo real. Essa experiência me ensinou a pensar sobre o processo pelo qual os dados são criados. Nesse caso, foi criado por um processo físico que eu poderia tocar. Na verdade, pude ver coisas que contribuíram para o porquê de os dados serem confusos ou por que não obtivemos um ponto de dados em uma determinada noite. Vejo paralelos diretos ao meu trabalho por vários anos em uma empresa de tecnologia que lidava com dados da web. Houve algum processo que gerou esses dados, e tive que pensar com cuidado sobre como gravamos esses dados e como o processo poderia correr bem ou mal. Essa experiência com dados do mundo real é a base da minha abordagem para desenvolver ferramentas de machine learning hoje em dia.

Outra competência que aprendi antes de ser cientista de dados foi me comunicar e ensinar. Fui professora universitária por vários anos e também desempenhei funções nas quais lidei com clientes. Nessas funções, pratiquei a competência de saber um conceito e de tentar transferir esse conhecimento ou entendimento para outra pessoa. Acredito firmemente que isso faz parte do papel da maioria dos cientistas de dados. Se treinarmos algum modelo ou fizermos alguma análise estatística, o valor disso é pequeno quando comparado a sermos capazes de pegar esse mesmo modelo ou análise e explicar o que significam, como funcionam ou como implementá-los em um contexto mais amplo.

## **Ao decidir se tornar cientista de dados, como você aprendeu novas competências?**

Acredito que os cursos universitários, bootcamps e materiais online são com certeza ótimas opções para pessoas diferentes em situações também diferentes. Como eu já tinha um doutorado, não queria voltar à universidade e investir mais dinheiro. Confesso que me candidatei a alguns bootcamps, e



eles não me aceitaram! Quando estava decidindo fazer essa transição de carreira para a ciência de dados, o que percebi foi que eu poderia fazer esse trabalho, mas necessitava demonstrar a outras pessoas que eu conseguia. De igual modo, precisava atualizar meu conhecimento sobre machine learning e algumas das técnicas, pois, quando eu estava na pós-graduação, o machine learning moderno não tinha de fato ingressado na astrofísica.

Segui a rota de cursos online e me dediquei muito a estudos autodidatas. Eu brinco que fiz todos os MOOCs (curso online aberto e massivo) que existiam: eram *muitos*. Levei aproximadamente seis meses mais ou menos sem muito o que fazer onde trabalhava e fiz vários cursos online. Eu estava fora da universidade há muito tempo e empolgada com o material. Até mesmo análise de dados eu estava sem fazer há um tempo, portanto, voltar para a análise de dados foi bem emocionante!

## **Você sabia que tipo de trabalho queria fazer ao entrar na ciência de dados?**

Quando pensei nas minhas opções a longo prazo e vi o que as pessoas estavam fazendo, como falavam sobre *analisar* x *construir* ciência de dados, me vi como alguém de análise. Me vi menos como engenheira e mais como cientista – uma pessoa que trabalha para entender eventos e responder a perguntas, mas não tanto construindo algo. Foi onde minha carreira começou. Eu era a única cientista de dados na Stack Overflow na maior parte do tempo em que lá estava em uma equipe com engenheiros de dados muito talentosos e com bastante conhecimento. Como era a única cientista de dados, meu trabalho envolvia um pouco de análise de dados e de construção de modelos. Agora, trabalhando em ferramentas de código aberto, tenho “engenheira de software” como nome do meu cargo e estou gastando mais energia na construção do que na análise.

## **O que você recomendaria às pessoas que desejam ser cientistas de dados?**

Algo que eu enfatizaria bastante é que você precisa demonstrar que pode fazer esse trabalho. Pode ser diferente para pessoas diferentes. Ainda é um campo jovem o suficiente e que as pessoas não têm certeza do que significa

ser um cientista de dados e quem pode ser um; ainda está muito indefinido. Ainda há muita incerteza a respeito do que se trata e os cargos são bem pagos, de forma que o risco percebido de uma empresa contratar alguém errado é muito alto, razão pela qual as empresas são muito avessas ao risco. As empresas precisam ter a certeza de que o candidato pode fazer esse trabalho. Algumas maneiras pelas quais vi pessoas demonstrarem saber fazer o trabalho é por meio de contribuições em código aberto, falando em meetups locais sobre projetos que fizeram e desenvolvendo um portfólio de projetos em um blog ou perfil do GitHub. Fiz todos os MOOCs e tudo o mais que precisava para aprender. Também comecei um blog sobre todos esses projetos. O que eu imaginei foi que esses projetos e posts do blog seriam algo que poderíamos falar em uma entrevista de emprego.

## **Resumo**

- Os quatro caminhos comprovados para aprender as competências para ser um cientista de dados são cursos universitários, bootcamps, aprendizagem no trabalho e também por conta própria.
- Cada um desses métodos tem vantagens e desvantagens em termos de materiais ensinados, duração e nível de automotivação necessário.
- Para escolher um caminho, reflita sobre quais competências você já tem, quais são seus pontos fortes e quais os seus recursos.

## CAPÍTULO 4

# Como montar um portfólio

Este capítulo abrange:

- A criação de um projeto atraente de ciência de dados
- Como iniciar um blog
- Procedimentos completos de projetos exemplares

Você agora terminou um bootcamp, um curso universitário, uma série de cursos online ou projetos de dados em seu trabalho. Parabéns! Está pronto para conseguir um emprego como cientista de dados. Certo?

Bem, talvez. A Parte II deste livro é toda sobre como encontrar, candidatar-se e conseguir um emprego em ciência de dados. Com certeza pode começar esse processo agora. Mas um outro passo pode realmente ajudá-lo a ser bem-sucedido: montar um portfólio. Um *portfólio* é um conjunto de projetos de ciência de dados que pode apresentar às pessoas para que elas possam ver que tipo de trabalho de ciência de dados você faz.

Um bom portfólio tem duas partes importantes: repositórios do GitHub (*repos* na sua abreviatura) e um blog. Um repo do GitHub hospeda o código para um projeto, enquanto o blog mostra suas competências de comunicação e a parte não codificada de seu trabalho de ciência de dados. A maioria das pessoas não quer ler milhares de linhas de código (seu repo); elas querem uma explicação rápida do que você fez e por que é importante (seu blog). E, quem sabe, até mesmo cientistas de dados do mundo inteiro podem ler seu blog, dependendo do tópico. Como vamos discutir na segunda parte deste capítulo, não precisa escrever posts de blog apenas sobre análises que fez ou modelos que construiu; também poderia explicar uma técnica estatística, escrever um tutorial para um método de análise de texto ou até mesmo compartilhar aconselhamento de carreira (como você escolheu seu curso universitário).

Não quer dizer que precisa ter um blog ou repos do GitHub cheio de projetos para ser um cientista de dados bem-sucedido. Na verdade, a maioria dos cientistas de dados não tem isso, e as pessoas conseguem empregos sem ter sempre um portfólio. Mas criar um portfólio é uma ótima maneira de ajudá-lo a se destacar e a praticar suas competências em ciência de dados e ficar ainda melhor. Esperamos que também seja divertido!

Este capítulo o orienta no processo de montar um bom portfólio. A primeira parte é sobre construir um projeto de ciência de dados e organizá-lo no GitHub. A segunda parte trata das melhores práticas para começar e compartilhar seu blog a fim de que aproveite ao máximo o trabalho feito. Depois, mostraremos dois projetos reais que fizemos para que você possa ver o processo do começo ao fim.

## 4.1 Como criar um projeto

Um projeto de ciência de dados começa com duas coisas: um conjunto interessante de dados e uma pergunta a fazer sobre o mesmo. Por exemplo, você pode pegar dados do censo do governo e perguntar: Como é que os dados demográficos de todo o país mudam ao longo do tempo?”. A combinação de pergunta e dados é a parte principal do projeto (Figura 4.1) e, com isso, é possível começar a fazer ciência de dados.

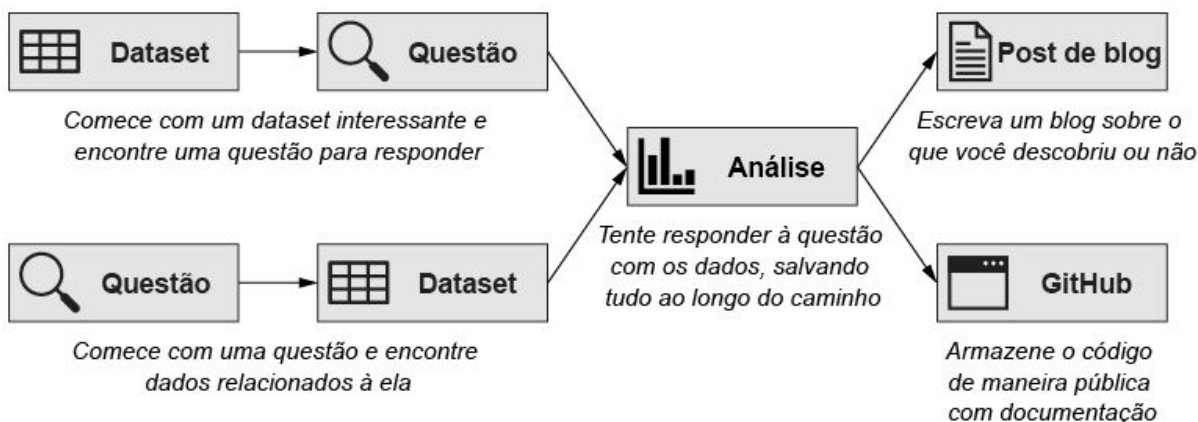


Figura 4.1 – Fluxo de criação de um projeto de ciência de dados.

### 4.1.1 Como encontrar dados e fazer uma pergunta

Quando estiver pensando em quais dados quer usar, o mais importante é

encontrar aqueles que sejam interessantes para você. Por que quer usar esses dados? Sua escolha de dados é uma forma de mostrar sua personalidade ou o conhecimento de domínio que tem da sua carreira ou de estudos anteriores. Se estiver em forma, por exemplo, pode investigar artigos sobre a Fashion Week e ver como os estilos mudaram nos últimos 20 anos. Se gostar de corridas, pode demonstrar como suas corridas mudaram ao longo do tempo e, talvez, ver se o tempo de corrida está relacionado com o clima.

Algo a não se fazer é usar o conjunto de dados Titanic, o MNIST ou qualquer outro conjunto de dados popular para iniciantes. Não é que essas experiências de aprendizagem não sejam boas; elas podem ser, mas provavelmente não encontrará nada novo que possa surpreender e intrigar os empregadores ou ensiná-los a ter mais conhecimento sobre você.

Às vezes, você deixa uma pergunta conduzi-lo ao seu conjunto de dados. Pode estar curioso, por exemplo, sobre como a distribuição de gênero das ênfases dos cursos universitários mudou com o tempo e se essa mudança está relacionada ao salário médio após a graduação. Então, recorreria ao Google e tentaria encontrar a melhor fonte desses dados.

Porém, talvez você não tenha uma pergunta específica para a qual esteja ansioso que as competências de ciência de dados encontrem respostas. Nesse caso, pode começar a procurar conjuntos de dados e ver se consegue fazer perguntas interessantes. A seguir, algumas sugestões por onde começar:

- *Kaggle.com* – Kaggle começou como um site para competições de ciência de dados. As empresas publicam um conjunto de dados e uma pergunta, oferecendo, normalmente, um prêmio à melhor resposta. Como as perguntas envolvem modelos de machine learning que tentam prever algo (ex., se alguém seria inadimplente em um empréstimo ou por quanto uma casa seria vendida), os usuários podem comparar modelos com base no desempenho em um conjunto de teste holdout e obter uma métrica de desempenho para cada um. O Kaggle também tem fóruns de discussão e “núcleos (kernels)” nos quais as pessoas compartilham seu código para você aprender como pessoas trataram o conjunto de dados. Como resultado, o Kaggle tem milhares de conjuntos

de dados com perguntas e exemplos de como outros as analisaram.

A maior vantagem do Kaggle é também seu maior inconveniente: por ter um conjunto de dados (geralmente limpo) e um problema, boa parte do trabalho já está feita. Você também tem milhares de pessoas lidando com o mesmo problema, portanto é difícil dar uma contribuição única. Uma forma de utilizar o Kaggle é pegar o conjunto de dados e fazer uma pergunta diferente ou uma análise exploratória. No geral, entretanto, pensamos que o Kaggle é melhor para aprender a mexer com um projeto e ver como você se saiu comparado aos demais, aprendendo, assim, o que os modelos deles tinham, em vez de ser parte do seu portfólio.

- *Conjunto de dados nos noticiários* – recentemente, muitas empresas de notícias começaram a deixar seus dados públicos. *FiveThirtyEight.com*, por exemplo, é um site que se concentra na análise das sondagens de opinião, política, economia e blogs de esporte, publicando dados que podem ser usados para artigos e até mesmo links para dados brutos diretamente do site do artigo. Embora esses conjuntos de dados em geral precisam de limpeza manual, o fato de estarem nos noticiários significa que uma pergunta óbvia provavelmente está associada a eles.
- *APIs* – APIs (interfaces de programação de aplicativos) são ferramentas de desenvolvedor que permitem acessar dados diretamente de empresas. Sabe quando você digita uma URL e chega a um site? As APIs são como URLs, mas, em vez de um site, você obtém dados. Alguns exemplos de empresas com APIs úteis são o *The New York Times* e o Yelp, os quais permitem acesso a seus artigos e avaliações, respectivamente. Algumas APIs até têm pacotes em linguagens R ou Python que facilitam especificamente o trabalho com elas. A *rtweet* para R, por exemplo, permite a obtenção dados do Twitter rapidamente para que possa encontrar tweets com uma hashtag específica, quais são os trending topics em Kyoto ou quais tweets Stephen King está favorecendo. Lembre-se de que existem limitações e termos de serviço para usar essas APIs. Neste momento, por exemplo, o Yelp limita você a cinco mil chamadas diárias para que você não consiga pegar todas as avaliações. As APIs são ótimas para fornecer dados extremamente robustos e organizados de várias fontes.

- *Dados abertos do governo* – muitos dados do governo estão disponíveis online. Você pode usar dados do censo, dados de emprego, pesquisa social geral e toneladas de dados do governo local, como as chamadas de emergência ou contagens de tráfego de uma cidade. Às vezes, você pode baixar esses dados diretamente como um arquivo CSV; outras vezes, precisa usar uma API. Nos EUA, é possível até enviar solicitações conforme a Lei de Liberdade de Informação a órgãos governamentais para obter dados que não estejam divulgados publicamente. As informações do governo são ótimas porque são frequentemente detalhadas e tratam de assuntos incomuns, como os dados sobre o registro de nomes de animais de estimação de uma cidade. A desvantagem das informações do governo é que muitas vezes não estão bem formatadas, como tabelas armazenadas em arquivos PDF.
- *Seus próprios dados* – existem muitos locais para baixar dados sobre você mesmo; os sites das redes sociais e serviços de email são duas grandes fontes. Se utilizar aplicativos para controlar sua atividade física, lista de leitura, orçamento, sono ou qualquer outra coisa, normalmente pode baixar esses dados também. Talvez possa construir um chatbot com base na troca de mensagens com seu cônjuge. Ou poderia ver as palavras mais comuns que usa em seus tweets e como elas mudaram ao longo do tempo. Talvez poderia acompanhar seu consumo de cafeína e exercício por um mês para ver se pode prever o quanto e quão bem você dorme. A vantagem de usar seus próprios dados é que seu projeto tem garantia de ser original: ninguém mais terá visto esses dados antes!
- *Web scraping* – coletar dados da web é uma maneira de extrair dados de sites que não têm uma API, essencialmente automatizando a visita em páginas e copiando os dados. Você pode criar um programa para pesquisar uma lista de 100 atores do site de um filme, carregar os perfis dos atores, copiar as listas de filmes em que eles estão e colocar esses dados em uma planilha. Porém, tem de ser cuidadoso: coletar dados de um site pode ser contra os termos de uso do site, e você pode ser banido. Você pode verificar o arquivo robots.txt de um site para descobrir o que é permitido. Você também quer ser legal com os sites: se acessar um deles muitas vezes, pode tirá-lo do ar. Supondo que os termos de serviço

permitam e que você dê tempo entre essas coletas, pode ser uma ótima maneira de conseguir dados originais.

O que torna um projeto paralelo interessante? Nossa recomendação é escolher uma análise exploratória na qual qualquer resultado provavelmente ensinará ao leitor algo ou demonstrará suas competências. Você pode criar um mapa interativo de chamadas emergenciais de uma cidade, codificado por categoria; esse mapa demonstra claramente suas competências de visualização e mostra, ainda, que pode escrever sobre os padrões que surgem. Por outro lado, se tentar prever o mercado de ações, provavelmente não será capaz, e é difícil um empregador avaliar suas competências se tiver um resultado negativo.

Outra dica é ver o que surge ao colocar sua pergunta no Google. Se os primeiros resultados forem artigos de jornal ou posts de blog que respondem exatamente à pergunta que estava fazendo, talvez seja melhor repensar sua abordagem. Às vezes, pode expandir a análise de outra pessoa ou trazer outros dados para agregar outra camada à análise, mas talvez seja necessário reiniciar o processo.

### **4.1.2 A escolha de uma direção**

Montar um portfólio não precisa demandar tanto tempo. A perfeição é definitivamente a inimiga aqui. É melhor ter algo do que nada. Os empregadores estão buscando acima de tudo alguma comprovação de que você pode programar e comunicar os dados. Você pode estar preocupado que as pessoas deem risada de seu código ou digam: “Uau, achávamos que essa pessoa poderia ser boa, mas olha que código horrível!”. É bem improvável que isso aconteça. Uma razão é que os empregadores ajustam as expectativas conforme o nível de experiência: não se espera que você programe como alguém da ciência da computação se for um cientista de dados em início de carreira. Geralmente, a maior preocupação é que não saiba programar nada.

Também é bom pensar sobre as áreas de ciências de dados que abordamos no Capítulo 1. Você pretende se especializar em visualização? Crie um gráfico interativo usando D3. Quer fazer processamento de linguagem natural? Utilize dados de texto. Machine learning? Faça predição de alguma



coisa.

Use seu projeto para forçar o aprendizado de algo novo. Fazer esse tipo de análise prática mostrará as lacunas do seu conhecimento. Quando os dados pelos quais realmente tem interesse encontram-se na web, você aprenderá a como fazer web scraping. Se achar que um gráfico em particular é feio, aprenderá a fazer visualizações melhores. Se estiver estudando por conta própria, fazer um projeto é uma boa maneira de superar a estagnação de não saber o que aprender em seguida.

Um problema comum com projetos automotivados é o *overscoping* (escopo excessivo). *Overscoping* quer dizer fazer tudo ou continuar adicionando mais coisas ao escopo. Você sempre pode seguir melhorando/editando/adicionando, mas assim nunca acaba. Uma estratégia é pensar como Hollywood e criar sequências. Você deve fazer uma pergunta e respondê-la, mas, se achar que pode querer revisita-la mais tarde, pode terminar sua pesquisa com uma pergunta ou tópico para investigação que requer mais buscas (ou mesmo “Continua no próximo capítulo...?”, se quiser).

Outra questão é não ser capaz de contornar problemas. Às vezes, os dados desejados não estão disponíveis. Ou não há quantidade suficiente. Ou não é possível limpá-los. Essas situações são frustrantes, e pode ser fácil desistir nesse momento. Mas vale a pena tentar descobrir como salvar o projeto. Você já fez trabalho suficiente para escrever um tutorial no blog, talvez sobre como coletou os dados? Os empregadores procuram pessoas que aprendem com seus erros e não têm medo de admiti-los. Mostrar o que deu errado para que outros possam evitar o mesmo destino também é valioso.

### **4.1.3 Preenchimento do README do GitHub**

Talvez você esteja em um bootcamp ou em um curso universitário no qual já esteja fazendo seus próprios projetos. Você até colocou seu código no GitHub. Será que basta?

Não! Um requisito mínimo para um repositório útil do GitHub é preencher o arquivo README. Para tanto, tem algumas perguntas para responder:

- *Qual é o projeto?* Quais dados ele usa? A qual pergunta ele está

respondendo? Qual foi o resultado: um modelo, um sistema de machine learning, um dashboard ou um relatório?

- *Como o repositório está organizado?* Essa pergunta implica, naturalmente, que o repo esteja de fato organizado de alguma forma! Há vários sistemas diferentes, mas um básico é dividir seu script em partes: como conseguir (se relevante) os dados, limpeza, exploração e análise final. Dessa forma, as pessoas sabem aonde ir, dependendo em que estão interessadas. Também sugere que manterá seu trabalho organizado quando for trabalhar para uma empresa. Uma empresa não quer arriscar contratá-lo e, quando for a hora de entregar um projeto, você entrega a alguém um script de cinco mil linhas não comentadas que pode ser impossível para eles entenderem e usarem. Uma boa gestão de projetos também o ajuda no futuro: se quiser reutilizar parte do código mais tarde, saberá aonde ir.

Mas embora seja bom fazer um projeto e torná-lo publicamente disponível em um repo documentado do GitHub, é muito difícil olhar para o código e entender por que ele é importante. Depois que você faz um projeto, a etapa seguinte é escrever um post no blog, para que as pessoas saibam por que aquilo que você fez foi legal e interessante. Ninguém se interessa por `análise_nomes_pets.R`, mas todos se interessem por “Usei linguagem R para encontrar os nomes de animais de estimação mais engraçados!”

## 4.2 Como iniciar um blog

Os blogs permitem que apresente seus pensamentos e projetos, mas também podem oferecer uma visão não técnica de seu trabalho. A gente sabe: você acabou de aprender todas essas coisas técnicas maravilhosas! Você quer se exibir! Mas ser um cientista de dados quase sempre implica comunicar seus resultados a uma audiência leiga, e um blog pode lhe dar experiência para traduzir seu processo de ciência de dados na linguagem dos negócios.

### 4.2.1 Tópicos em potencial

Suponha que tenha criado um blog. Será que as pessoas ficarão realmente interessadas nos seus projetos? Você nem sequer tem o título de cientista de

dados ainda; como pode ensinar alguma coisa?

Algo bom para recordar é que se trata de um bom momento para ensinar quem está poucas etapas atrás de você. Logo após ter aprendido um conceito (por exemplo, usar integração contínua para seu pacote ou fazer um modelo TensorFlow), você ainda entende os equívocos e as frustrações que teve. Anos mais tarde, é difícil pensar como principiante. Você já teve um professor que claramente era muito inteligente e, no entanto, não conseguia comunicar os conceitos? Não duvidou que eles conheciam o tópico, mas não conseguiam explicar e pareciam frustrados porque você não conseguia entender logo.

Tente pensar no seu público como sendo você seis meses atrás. O que aprendeu desde então? Quais recursos gostaria que estivessem disponíveis? Esse exercício é ótimo também para comemorar seu progresso. Com tanta coisa para aprender em ciência dos dados, é fácil sentir que nunca fez o suficiente; é bom fazer uma pausa para ver o que já conseguiu fazer.

Você pode agrupar posts de blog de ciência de dados em quatro categorias:

- *Tutoriais pesados de código* – os tutoriais mostram aos seus leitores como fazer coisas como web scraping ou deep learning em Python. Seus leitores serão em geral outros aspirantes ou praticantes de ciência de dados. Embora chamemos de tutoriais *pesados* de código, você normalmente ainda vai querer que haja o máximo possível de linhas de texto como código, se não mais. O código via de regra não é autoexplicativo; você precisa pegar o leitor pela mão e explicar o que cada parte faz, por que quer fazer isso e quais são os resultados.
- *Tutoriais pesados em teoria* – esses tutoriais ensinam aos leitores um conceito estatístico ou matemático, por exemplo: o que é Bayes empírico ou como a análise de componentes principais funciona. Eles podem ter algumas equações ou simulações. Como acontece com os tutoriais pesados em código, seu público normalmente é de outros cientistas de dados, mas você deve escrever de forma que qualquer pessoa que tenha alguma base matemática possa acompanhar. Os tutoriais pesados em teoria são especialmente bons para demonstrar suas competências de comunicação; há um estereótipo que pessoas muito

técnicas, sobretudo as que têm doutorado, não conseguem explicar bem os conceitos.

- *Um projeto divertido que você fez* – como esperamos tê-lo convencido na Seção 4.1, você não precisa trabalhar apenas com inovações no reconhecimento de imagens médicas. Também é possível descobrir qual dos filmes da saga *Crepúsculo* usou apenas palavras da peça *A Tempestade* de Shakespeare. Julia Silge, por exemplo, que entrevistamos no Capítulo 3, usou redes neurais para gerar texto que soava como a Jane Austen. Esses posts de blog podem se concentrar mais nos resultados ou no processo, dependendo da parte mais interessante do projeto.
- *Escrevendo sobre sua experiência* – você não precisa apenas escrever tutoriais ou a respeito de seus projetos de ciência de dados. Pode falar sobre sua experiência em uma reunião ou conferência de ciência de dados: quais palestras achou interessante, conselho para pessoas indo a uma primeira conferência ou alguns recursos que os palestrantes compartilharam. Esse tipo de post pode ser útil para pessoas que estão considerando participar do mesmo evento no ano seguinte ou que não podem participar de conferências por questões logísticas ou financeiras. Novamente, esses tipos de posts de blog dão aos empregadores uma ideia sobre como você pensa e se comunica.

### 4.2.2 Logística

Onde escrever? Para um blog, você tem duas opções importantes:

- *Fazer seu próprio site*. Se trabalhar em linguagem R, sugerimos usar o pacote *blogdown*, que permite criar um site para um blog usando o código R (loucura, né?). Se usar Python, Hugo e Jekyll são duas opções, e ambas permitem criar sites estáticos de blog, além de ter vários temas que outras pessoas fizeram, permitindo que escreva posts de blog no editor markdown. Nossa sugestão é que não se preocupe muito com seu tema e estilo; basta escolher um que o agrade. Não tem nada pior do que não escrever posts de blog porque se distraiu mudando a aparência do blog. Escolher algo simples certamente é o melhor; pode ser complicado

mudar o tema, por isso é prudente não escolher um que pode cansá-lo em seis meses.

- *Usar o Medium ou outra plataforma de blog.* Medium é uma plataforma de publicação online e gratuita. A empresa geralmente não cria conteúdo; em vez disso, hospeda conteúdo para centenas de milhares de autores. Medium e sites como esse são boas opções se quiser começar rapidamente, pois não tem que se preocupar sobre hospedagem ou começar um site; tudo que você precisa fazer é clicar “novo post”, começar a escrever e publicar. Você também pode conseguir mais tráfego quando as pessoas pesquisam no site de blogs termos como *ciência de dados* ou *Python*. Mas uma preocupação é que você esteja à mercê da plataforma. Se a empresa alterar o modelo de negócio e começar a cobrar algum valor de acesso, por exemplo, não há nada que se possa fazer para manter seus posts de blog gratuitos. Você também não consegue criar uma seção de biografia real ou adicionar outros conteúdos, como uma página com links para palestras que ministrou.

Uma pergunta comum sobre blogs é com que frequência é bom postar e que tamanho os posts devem ter. Definitivamente são escolhas pessoais. Vimos pessoas com microblogs publicando textos curtos várias vezes por semana. Outras levam meses entre um texto e outro, publicando artigos mais longos. Há algumas limitações. É preciso evitar que textos não se pareçam com um livro como *Ulisses*. Se seu texto for muito longo, pode dividi-lo em partes. Mas é bom mostrar que consegue se comunicar de forma concisa, já que essa é uma das principais competências da ciência de dados. Os executivos e até mesmo seu gerente provavelmente não querem ou não precisam saber de todas as suas tentativas e falhas. Embora possa optar por incluir um breve resumo daquilo que pode ter dado errado, é preciso chegar ao ponto e ao caminho final rapidamente. Uma exceção, entretanto, é se seu método final for surpreender leitores. Se não tiver usado a biblioteca mais popular para um problema, por exemplo, talvez queira explicar que não usou porque descobriu que a biblioteca não funcionava.

E se estiver preocupado que ninguém lerá seu blog e que todo seu trabalho será para nada? Bem, uma razão para ter um blog é que ele o ajuda a se candidatar a empregos. Você pode colocar links para textos do seu blog

no seu currículo quando faz referência a projetos de ciência de dados e até mesmo mostrá-lo em entrevistas, especialmente se os textos têm uma visualização ou dashboards interativos e agradáveis aos olhos. Não é importante ter centenas ou milhares de leitores. Pode ser bom se você receber palmas no Medium ou ser incluído na newsletter de uma empresa de ciência de dados, mas é mais importante ter um público que leia, valorize e envolva-se com o material do que ter métricas altas.

Não quer dizer que não haja o que fazer para aumentar o número de leitores. Por um lado, você deve fazer propaganda. Embora seja um clichê, ter uma #marca é útil para construir uma rede a longo prazo. Mesmo que algo pareça simples, provavelmente é novo para praticantes de ciência de dados, pois o campo é muito grande. As pessoas nas empresas em que você quer trabalhar podem até mesmo ler seu material! O Twitter é um bom lugar para começar. Você pode compartilhar quando liberar um texto e usar as hashtags apropriadas para conseguir mais leitores.

Mas seu blog é valioso mesmo se ninguém o ler (além de seu companheiro/a e animal de estimação). Escrever um texto de blog é uma boa prática, pois o força a estruturar seus pensamentos. Assim como ensinar pessoalmente também o ajuda a perceber quando não sabe tão bem assim quanto achava que sabia.

## **4.3 Trabalhar em projetos exemplares**

Nesta seção, demonstraremos dois projetos exemplares, da ideia inicial à análise até um produto público final. Usaremos projetos reais que as autoras deste livro fizeram: criar um aplicativo da web para que freelancers de ciência de dados encontrassem bons trabalhos e aprender redes neurais treinando uma em um conjunto de dados de placas de carros banidas.

### **4.3.1 Freelancers de ciência de dados**

*Emily Robinson*

#### **A pergunta**

Quando eu era uma aspirante à cientista de dados, interessei-me por uma

maneira com a qual alguns cientistas de dados ganhavam dinheiro extra: fazendo freela. *Fazer freela* é fazer projetos para alguém com quem você não tem um vínculo empregatício, seja essa outra pessoa física ou uma grande empresa. Esses projetos variam de poucas horas a meses de trabalho em tempo integral. É possível encontrar muitos trabalhos de freelancer em sites como o UpWork, mas como a ciência de dados é um campo muito vasto, os trabalhos nessa categoria podem ser qualquer coisa, desde desenvolvimento web, passando por uma análise no Excel, até processamento de linguagem natural em terabytes de dados. Decidi ver se eu conseguia ajudar freelancers a vasculharem milhares de ofertas de trabalho para encontrar aquelas que são as mais adequadas para eles.

## **A análise**

Para coletar os dados, usei a API do UpWork para obter os trabalhos disponíveis no momento e os perfis de todos na categoria de Ciência de Dados e Análise de Dados. Eram 93 mil freelancers e três mil trabalhos. Embora a API tenha facilitado bastante o acesso aos dados (pois não precisei fazer web scraping), ainda tive que fazer funções para realizar centenas de chamadas da API, lidar com as falhas dessas chamadas da API e transformar os dados para poder usá-los. Mas a vantagem desse processo foi que, como os dados não estavam prontamente disponíveis, não havia centenas de outras pessoas trabalhando no mesmo projeto, como teria havido se eu usasse dados de uma competição do Kaggle.

Depois de ter os dados no formato correto, fiz uma análise exploratória. Analisei como os níveis de educação e o país afetavam o quanto os freelancers ganhavam. Também construí um gráfico da correlação das competências que freelancers listaram, o que mostrou os tipos diferentes de freelancers: desenvolvedores web (PHP, jQuery, HTML e CSS), financeiro e contabilidade (contabilidade financeira, contabilidade e análise financeira) e coleta de dados (entrada de dados, geração de leads, mineração de dados e web scraping) com o conjunto “tradicional” de competências de ciência de dados (Python, machine learning, estatística e análise de dados).

Por fim, criei uma pontuação de similaridade entre o texto do perfil e o texto do trabalho, e combinei essa pontuação com a sobreposição de

competências (as competências listadas de freelancers e do trabalho) para criar uma pontuação relacionando um freelancer a um trabalho.

## **O produto final**

Neste caso, acabei não escrevendo um texto no blog. Em vez disso, fiz um aplicativo web interativo no qual alguém poderia colocar o texto do perfil, competências e requisitos para os trabalhos (como pontuação mínima de feedback do empregador do trabalho e quanto tempo o trabalho levaria), com os trabalhos disponíveis filtrados para atender a esses requisitos e na ordem em que se encaixavam melhor ao usuário.

Não deixei que a perfeição fosse a inimiga aqui. Há muitas maneiras que eu poderia ter feito o projeto melhor. Puxei os trabalhos uma vez só, e, como fiz esse projeto há quatro anos, o aplicativo ainda funciona, mas nenhum dos trabalhos está disponível agora. Para tornar o aplicativo valioso a longo prazo, eu precisaria puxar os trabalhos todas as noites e atualizar as listas. Também poderia ter feito um algoritmo de correspondência mais sofisticado, ter acelerado o tempo de carregamento inicial do app e ter deixado a aparência mais bonita. Apesar dessas limitações, o projeto cumpriu alguns objetivos importantes. Mostrou que consegui fazer um projeto e permitir que outras pessoas interagissem com ele, em vez de se limitar às análises estáticas que ficariam em meu notebook. Foi um caso de uso no mundo real: ajudar freelancers a encontrar trabalhos. E passei pelo ciclo completo de um projeto de ciência de dados: coletar dados, limpá-los, executar análises exploratórias e gerar um produto final.

### **4.3.2 Treinar uma rede neural em placas de carro ofensivas**

*Jacqueline Nolis*

#### **A pergunta**

Conforme fui crescendo como cientista de dados, frustrava-me quando via textos hilários de blog em que as pessoas treinavam redes neurais para gerar coisas como novos nomes de banda, novos Pokémons e receitas culinárias bizarras. Achava esses projetos ótimos, mas não sabia como fazê-los por conta própria! Um dia, lembrei-me de que havia ouvido falar de um



conjunto de dados de todas as placas de carros personalizadas que foram rejeitadas pelo estado do Arizona, EUA, por serem muito ofensivas. Se eu pudesse obter esse conjunto de dados, seria perfeito para finalmente aprender a como fazer redes neurais – eu poderia fazer minhas próprias placas de carro ofensivas (Figura 4.2)!



*Figura 4.2 – Resultado da amostra de rede neural do gerador de placa de carro ofensiva<sup>1</sup>.*

## **A análise**

Depois de enviar uma solicitação de registros públicos para o Departamento de Transportes do Arizona, recebi uma lista de milhares de placas ofensivas. Eu não sabia nada sobre redes neurais, então, depois de receber os dados, comecei a vasculhar na internet procurando por textos de blog que descrevessem como fazer uma. Como usuária da linguagem R, fiquei feliz por encontrar o pacote Keras da RStudio para fazer redes neurais em R.

Carreguei os dados em R e conferi o exemplo do pacote Keras da RStudio para gerar texto com redes neurais. Alterei o código para trabalhar com os dados das placas; o exemplo da RStudio era gerar sequências de texto longo, mas eu queria treinar em placas de carro de sete caracteres. Eu tive que criar vários pontos de dados de treinamento para o meu modelo a partir de cada placa de carro (um ponto de dados para predizer cada caractere na placa).

Em seguida, treinei o modelo da rede neural, embora não tenha inicialmente sido útil. Depois de pausar o projeto por um mês, voltei e percebi que meus dados não estavam sendo processados corretamente.

Quando eu corriji esse problema, os resultados que a rede neural gerou foram fantásticos. Por fim, embora eu não tenha mudado muito o exemplo da RStudio, senti-me muito mais confiante na criação e no uso de redes neurais.

## O produto final

Escrevi um texto no blog sobre o projeto que explica como consegui os dados, o ato de processá-los para estarem prontos para a rede neural e como modifiquei o código de exemplo da RStudio para que trabalhasse para mim. O texto do blog era algo mais no estilo: “Sou nova em redes neurais e aqui está o que aprendi”; não fingi que já sabia como tudo isso funcionava. Como parte do post do blog, fiz uma imagem que tirava o resultado de texto do meu modelo neural e fiz com que se parecesse com as placas de carro do Arizona. Também coloquei o código no GitHub.

Desde que escrevi esse post no blog e disponibilizei meu código, muitas pessoas o modificaram para fazer suas próprias redes neurais engraçadas. O que aprendi com esse projeto divertido acabou me ajudando a fazer modelos de machine learning de alto impacto para consultorias importantes. Só porque o trabalho original não é sério não significa que ele não tenha valor!

## 4.4 Entrevista com David Robinson, cientista de dados

David Robinson é o coautor (com Julia Silge) do pacote *tidytext* em linguagem R e do livro da O'Reilly *Text Mining with R* (Mineração de texto com R, em tradução livre). Ele também é o autor do ebook autopublicado *Introduction to Empirical Bayes: Examples from Baseball Statistics* (Introdução a Bayes Empírico: exemplos de estatísticas de baseball, em tradução livre) e dos pacotes “broom” e “fuzzyjoin” em R. É doutor em biologia quantitativa e computacional pela Universidade de Princeton. Robinson escreve sobre estatística, análise de dados, educação e programação em R em seu famoso blog: [varianceexplained.org](http://varianceexplained.org).

## **Como começou o blog?**

Comecei primeiramente o blog quando estava me candidatando a empregos perto do fim do meu doutorado, porque me dei conta de que eu não tinha muita coisa na internet que mostrasse minhas competências de programação ou estatística. Quando lancei meu blog, lembro de ter medo de que depois que escrevesse alguns textos que já tinha prontos na cabeça, ficaria sem ideias. Mas fiquei surpreso ao descobrir que continuei a ter novas ideias sobre as quais queria escrever: conjuntos de dados que gostaria de analisar, opiniões que queria compartilhar e métodos que queria ensinar. Tenho escrito no blog com moderação e consistência há quatro anos desde então.

## **Há oportunidades específicas que tenha aproveitado do trabalho público?**

Conseguí meu primeiro emprego com algo que escrevi publicamente online. A Stack Overflow entrou em contato comigo por causa de uma resposta que escrevi sobre o site de estatísticas deles. Eu havia escrito aquela resposta há anos, mas alguns engenheiros a encontraram e ficaram impressionados com ela. Essa experiência realmente me levou a acreditar firmemente na produção de produtos públicos, pois às vezes os benefícios aparecerão meses ou anos depois e podem levar a oportunidades nunca esperadas.

## **Há pessoas que você acha que se beneficiariam em especial em fazer trabalho público?**

As pessoas cujos currículos não mostram suas competências em ciências de dados e que não tenham uma formação típica, como ter doutorado ou experiência como analista de dados, se beneficiariam em especial do trabalho público. Quando estou avaliando um candidato, se ele não tiver essas credenciais, é difícil dizer se será capaz de fazer o trabalho. Mas minha maneira favorita de avaliar um candidato é ler uma análise por ele publicada online. Se eu puder ver alguns gráficos criados por ele, como explicou a história e como lidou com os dados, posso começar a entender se poderá ser bom na função.

## **Sua opinião sobre o valor do trabalho público mudou com o tempo?**

Eu via projetos como algo que consistia em progredir constantemente enquanto estivesse trabalhando em algo. No período de pós-graduação, uma ideia não valia muito a pena, mas depois tornou-se um código, um rascunho, um rascunho finalizado e, então, um artigo publicado. Pensei, na época, que meu trabalho ficava cada vez mais valioso.

Desde então, percebi que estava pensando nisso de uma maneira completamente errada. Qualquer coisa que ainda está no computador, completa ou não, é inútil. Se não estiver no mundo, foi desperdiçada até o momento, e qualquer coisa que está no mundo é muito mais valiosa. O que me fez perceber foram alguns artigos que redigi no período de pós-graduação e que nunca publiquei. Eles deram bastante trabalho, mas continuei sentindo que não estavam exatamente prontos. Anos mais tarde, acabei esquecendo sobre o que eram, e eles não agregaram nada ao mundo. Se eu tivesse escrito alguns textos no blog, feito alguns tweets e um pacote aberto bem simples de código aberto, tudo isso teria agregado algum valor.

## **Como tem ideias para seus textos de análise de dados?**

Criei o hábito de que toda vez que me deparo com um conjunto de dados, faço o download e o analiso rapidamente, executando algumas linhas de código para ter uma noção dos dados. Isso ajuda a criar certo gosto por ciência de dados, e trabalhar em projetos suficientes faz com que já tenha uma ideia melhor sobre quais dados será interessante escrever e quais seriam melhores para deixar de lado.

Meu conselho é que sempre que tiver a oportunidade de analisar dados, mesmo que não esteja no seu trabalho do momento ou se achar que pode não ser interessante para você, dê uma olhada e veja o que pode encontrar em apenas alguns minutos. Escolha um conjunto de dados, decida um período determinado de tempo, faça todas as análises que puder e publique. Pode não ser um post perfeito ou que não encontre tudo o que esperava para responder a todas as perguntas que queria. Mas, ao definir o objetivo de que um conjunto de dados vire um post, você pode começar a adquirir esse hábito.

## Qual é seu último conselho para os cientistas de dados júnior e aspirantes?

Não se estressem em acompanhar a tecnologia de ponta da área. É tentador, quando se inicia na área de ciência de dados e machine learning, pensar que deve começar a trabalhar com deep learning ou outros métodos avançados. Porém, lembre-se de que esses métodos foram desenvolvidos para resolver alguns dos problemas mais difíceis na área. Não são necessariamente os problemas que enfrentará como cientista de dados, especialmente no início da carreira. Você deve começar se sentindo bastante confortável em transformar e visualizar dados, em programar com uma grande variedade de pacotes e em usar técnicas estatísticas, como testes de hipótese, classificação e regressão. Vale a pena entender esses conceitos e conseguir aplicá-los antes de começar a preocupar-se com conceitos mais modernos.

## Resumo

- Ter um portfólio de projetos de ciência de dados compartilhados em um repo do GitHub e ter um blog pode ajudá-lo a conseguir um emprego.
- Há muitos lugares para encontrar bons conjuntos de dados para um projeto paralelo. O mais importante é escolher algo que seja interessante para você e um pouco incomum.
- Você não precisa escrever em um blog somente sobre seus projetos paralelos. Você também pode compartilhar tutoriais ou sua experiência em um bootcamp, conferência ou curso online.

## Recursos dos capítulos 1–4

### Livros

*Practical Data Science with R*, 2ª ed., de Nina Zumel e John Mount (Manning Publications)

Este livro é uma introdução à ciência de dados que usa a linguagem R como a ferramenta principal. É um complemento excelente deste livro que você está lendo, porque vai muito além dos componentes técnicos do trabalho. Ele trata de conjuntos de dados, pensando sobre as perguntas que pode fazer sobre eles e como fazê-lo, interpretando depois os resultados.

*Doing Data Science: Straight Talk from the Frontline*, de Cathy O’Neil e Rachel Schutt (O’Reilly Publications)

Outra introdução à ciência de dados, este livro é uma mistura de teoria e prática. Tem uma visão ampla do campo e tenta abordá-lo de vários ângulos em vez de ser um conjunto de estudos de caso.

*R for Everyone*, 2ª ed., de Jared Lander, e *Pandas for Everyone*, de Daniel Chen (Addison-Wesley Data and Analytics)

*R for Everyone* e *Pandas for Everyone* são dois livros da série da Addison-Wesley Data and Analytics. Eles tratam do uso das linguagens R e Python (via pandas), desde funções básicas até análises avançadas e resolução de problemas de ciência de dados. Para pessoas que sentem a necessidade de ajuda para aprender qualquer um desses tópicos, são ótimos recursos.

*Think Like a Data Scientist: Tackle the Data Science Process Step-by-Step*, de Brian Godset (Manning Publications)



*Think Like a Data Scientist* é um livro introdutório de ciência de dados estruturado em torno de como o trabalho da ciência dos dados é realmente feito. Trata da definição do problema e da criação do plano, resolvendo problemas de ciência de dados e apresentando suas descobertas aos outros. Este livro é melhor para pessoas que entendem os conceitos técnicos da ciência de dados, mas são novas em projetos de longo prazo.

*Getting What You Came For: The Smart Student's Guide to Earning an M.A. or a Ph.D.*, de Robert L. Peters (Farrar, Straus and Giroux)

Se você decidiu ingressar na universidade para fazer um mestrado ou doutorado, embarcará em uma viagem longa e cansativa. Entender como ser aprovado em exames e qualificações, perseverar na pesquisa e terminar rapidamente não são coisas exatamente ensinadas. Embora este livro seja bastante antigo, as lições que ensina sobre como obter sucesso aplicam-se à pós-graduação ainda hoje.

*Bird by Bird: Some Instructions on Writing and Life*, de Anne Lamott  
(Anchor)

*Bird by Bird* não é somente um guia para escrever, mas também para a vida. O título surgiu de algo que o pai de Anne Lamott disse ao irmão dela quando estava desesperado fazendo um relatório sobre pássaros que ele tivera três meses para fazer, mas deixara para a última noite: “Pássaro a pássaro, cara. Faça pássaro a pássaro”. Se você tem sentido dificuldade com o perfeccionismo ou em pensar sobre o que escrever, esse livro pode ser bom.

## **Textos de blog**

*Bootcamp rankings*, de Switchup.org

<https://www.switchup.org/rankings/best-data-science-bootcamps>

Switchup tem uma lista dos 20 principais bootcamps com base em comentários de alunos. Embora você possa achar que as avaliações não sejam corretas, este blog ainda é um bom ponto de partida para escolher a quais bootcamps se candidatar.

*What's the Difference between Data Science, Machine Learning, and Artificial Intelligence?*, de David Robinson

<http://varianceexplained.org/r/ds-ml-ai>

Se estiver confuso sobre o que é ciência de dados x machine learning x inteligência artificial, este texto esclarece de maneira produtiva. Embora não haja definições universalmente aceitas, gostamos dessa taxonomia na qual a ciência de dados produz percepções, o machine learning gera previsões e a inteligência artificial leva a ações.

*What You Need to Know before Considering a PhD*, de Rachel Thomas

<https://www.fast.ai/2018/08/27/grad-school>

Se estiver pensando que precisa de um doutorado para ser um cientista de dados, leia este blog primeiro. Thomas mostra os custos significativos de fazer um doutorado (em termos de custos potenciais de saúde mental e de oportunidade de carreira) e desfaz o mito de que é necessário um doutorado para fazer investigação de ponta em deep learning.

*Thinking of Blogging about Data Science? Here Are Some Tips and Possible Benefits*, de Derrick Mwiti

<http://mng.bz/gVEx>

Se o Capítulo 4 não o convenceu dos benefícios de fazer um blog, talvez este texto o convença. Mwiti também dá algumas dicas excelentes sobre como tornar seus posts interessantes, incluindo usar listas e novos conjuntos de dados.

*How to Build a Data Science Portfolio*, de Michael Galarnyk

<http://mng.bz/eDWP>

Este é um texto excelente e detalhado sobre como fazer um portfólio de ciência de dados. Galarnyk mostra não só quais tipos de projetos incluir (e não incluir) em um portfólio, mas também como incorporá-los em seu currículo e compartilhá-los.

- 
- <sup>1</sup> N.T.: As placas trazem palavras que poderiam ser consideradas ofensivas em inglês: 1. Blazen significa o estado em que a pessoa fica depois de fumar maconha; 2. Xtassy vem da droga ecstasy; 3. H8terr é a palavra hater, uma pessoa que odeia/critica alguma coisa; 4. Big4ch soa como "bitch", uma palavra pejorativa que significa "vadia".



## PARTE II

# Como encontrar empregos em ciência de dados

Agora que você está preparado para conseguir um trabalho de ciência de dados, é hora de correr atrás. Esta parte do livro aborda tudo o que precisa saber para realizar uma busca de emprego bem-sucedida, começando por encontrar vagas abertas e terminando com a negociação e aceitação de uma oferta de emprego. O processo de busca de emprego em ciência de dados tem algumas peculiaridades por causa da natureza da área. Prepararemos você para analisar as muitas vagas de empregos de *cientista de dados* e o que as empresas estão procurando em um estudo de caso real. Embora essa parte seja especialmente útil se não tiver trabalhado com ciência de dados antes, o conteúdo ainda pode ser útil como uma atualização para cientistas de dados com menos ou mais experiência.

O Capítulo 5 aborda a busca por vagas em ciência de dados e como lidar com a imensa variedade de vagas de emprego. O Capítulo 6 ensina a criar um currículo de ciência de dados e elaborar uma carta de apresentação, dando exemplos para basear seus documentos e os princípios por trás deles. O Capítulo 7 é sobre o que esperar e como se preparar para a entrevista de emprego em ciência de dados, desde a entrevista por telefone até a reunião final na empresa. O Capítulo 8 explica o que dizer quando recebe uma oferta de uma empresa, incluindo como decidir se a aceita, além de por que e como negociar.

## CAPÍTULO 5

# **A busca: como identificar o emprego certo para você**

Este capítulo abrange:

- Como encontrar vagas abertas que possam ser boas
- Como decodificar as descrições de emprego para entender como as funções realmente são
- Como escolher os empregos para os quais pretende candidatar

Você tem as competências e o portfólio. Tudo o que está faltando é o emprego de ciência de dados! No entanto, é provável que o processo de busca de emprego leve algum tempo. Mesmo aqueles processos seletivos bem-sucedidos levam pelo menos um mês entre a candidatura e o recebimento de uma oferta, na verdade, geralmente levam vários meses. Ao definir algumas práticas recomendadas neste capítulo, esperamos tornar o processo o menos difícil possível.

Neste capítulo, focamos em como buscar vagas de ciência de dados. Em primeiro lugar, abordamos todos os locais onde pode buscar essas vagas, a fim de garantir que, sem saber, você não restrinja suas opções. Em seguida, mostramos como decodificar as descrições para descobrir quais competências são realmente necessárias (spoiler: nem todas elas) e como as vagas podem ser. Por fim, mostramos como escolher aquelas mais adequadas para você, utilizando o conhecimento aprendido sobre as competências de ciência de dados e os modelos de empresa nos primeiros quatro capítulos.

## 5.1 Como encontrar vagas

Antes de se preocupar com a criação de um currículo “perfeito” e com a carta de apresentação bem feita, precisa saber para onde enviá-los! Sites de vagas como LinkedIn, Indeed e Glassdoor são bons lugares para começar sua busca. Vale a pena olhar em mais de um site, pois nem todas as empresas publicarão em cada um deles. Se fizer parte de um grupo com pouca representação na tecnologia, procure também sites específicos que ofereçam vagas para você, como o POCIT e o Tech Ladies, voltados a negros e a mulheres na tecnologia, respectivamente. O tipo de vaga para o qual está se candidatando também pode influenciar o local da busca; por exemplo, há sites de vagas para tipos específicos de empresas, como startups (AngelList) e de tecnologia (Dice).

Certifique-se de fazer uma busca ampla. Conforme discutido no Capítulo 1, as vagas em ciência de dados têm muitos outros nomes além de cientista de dados. Empresas diferentes usam nomes diferentes para funções similares, e algumas estão até mesmo mudando o que os nomes dos cargos significam. Então, todas as pessoas que eram analistas de dados em um certo momento podem virar cientistas de dados no ano seguinte, sem mudanças de responsabilidade!

Alguns exemplos de nomes que podem ser encontrados (Figura 5.1) são:

- *Analista de dados* – este cargo é muitas vezes uma posição júnior e pode ser uma ótima maneira de começar no campo se não tiver uma formação de ciência, tecnologia, engenharia ou matemática (CTEM) e também não tiver feito análise de dados para uma empresa antes. Conforme discutimos mais adiante na Seção 5.1.1, é bom ter um cuidado extra com uma posição de analista de dados para garantir que a função envolverá programação e estatística ou machine learning.
- *Quantitativos, de produto, de pesquisa ou outros tipos de analistas* – estas funções têm ainda mais diversidade do que os analistas de dados em termos de responsabilidades. Você pode fazer exatamente o mesmo tipo de trabalho que os cientistas de dados de outras empresas ou passar os dias com planilhas antigas do Microsoft Excel.
- *Engenheiro de machine learning* – conforme está implícito no título, são

empregos que se concentram na parte de machine learning da ciência de dados e geralmente exigem uma boa formação em engenharia. Se é graduado em ciência da computação ou trabalha como engenheiro de software, esta função pode ser excelente para você.

- *Cientista de pesquisa* – estes cargos muitas vezes exigem doutorado, embora possa haver espaço para negociação se tiver um mestrado em ciência da computação, estatística ou em um campo estreitamente relacionado.

Quando estiver iniciando a sua busca, tente pesquisar simplesmente *dados* em um desses sites de vagas e passe em torno de uma hora lendo as vagas oferecidas. Você terá uma ideia melhor de quais setores são representados em sua área e quais tipos de vagas estão abertas. Reconhecerá padrões que permitirão que leia mais rapidamente novas vagas. Encontrar vagas que sejam boas para você, em vez de todas as vagas disponíveis, limitará a área a um número gerenciável. Não se preocupe muito com o título de uma vaga. Use a descrição apresentada para avaliá-la.



*Figura 5.1 – Alguns títulos de emprego que não incluem “ciência de dados” que podem ser encontrados ao fazer uma pesquisa.*

Seja extremamente cauteloso sobre pensar a procura de emprego como um jogo de números. Se estiver buscando em uma grande cidade tecnológica ou em várias cidades, encontrará centenas de vagas. Conferir sites de vagas pode tornar-se rapidamente uma obsessão porque é uma forma fácil de se sentir produtivo (“Hoje eu li 70 descrições de vagas!”). Como com o Twitter e o Facebook, verificar constantemente atualizações pode ser viciante. Verificar mais do que a cada três a cinco dias via de regra não compensa. Verificar apenas uma vez por mês pode significar perder uma boa oportunidade, mas nenhuma empresa preenche uma vaga (que esteja de

fato aberta) no prazo de dois dias após a publicação num site de emprego.

Se estiver interessado em empresas específicas, consulte a seção “trabalhe conosco” dos sites delas. Assim como pesquisaria vários títulos de vaga, verifique também diferentes departamentos. Algumas empresas podem colocar a ciência de dados em departamentos financeiros, de engenharia ou outros, por isso, se não diversificar sua pesquisa, pode ser que não encontre essas vagas.

**RECÉM-FORMADOS** Quando estiver buscando emprego, procure vagas especificamente intituladas “recém-formado”, “júnior”, “auxiliar” e “início de carreira”. Procure também centros de promoção de vagas da sua universidade para receber ajuda e participe das feiras de trabalho no campus.

### 5.1.1 Como decodificar as descrições

Ao ler as descrições das vagas, aquelas de ciência de dados parecem cair em uma de duas categorias:

- *Uma vaga de analista de inteligência de negócios* – nesta função, usará ferramentas de inteligência de negócios, como o Excel e o Tableau, talvez um pouco de SQL, mas, no geral, você não programará. Se quiser melhorar suas competências de programação, de machine learning, de estatística e seu conhecimento de engenharia de dados, tais vagas não são a melhor opção.
- *Um unicórnio* – no outro extremo, a vaga requer alguém com doutorado em ciência da computação e que também trabalhe como cientista de dados há mais de cinco anos; que seja um especialista em estatística de ponta, deep learning (aprendizado profundo) e na comunicação com parceiros de negócios; que tenha experiência com diversas responsabilidades, desde machine learning no nível de produção a criar dashboards para executar testes A/B. Esses tipos de descrições geralmente significam que a empresa não sabe o que está buscando e espera que um cientista de dados resolva todos os problemas sem nenhum suporte.

No entanto, não se preocupe: garantimos que há mais do que esses dois tipos de vagas. Uma forma melhor de pensar nessas vagas é em termos de experiência. A empresa está buscando alguém para formar um

departamento próprio e não tem nenhuma infraestrutura de pipeline de dados em vigor? Ou busca mais alguém para a equipe de ciência de dados atualmente produtiva, esperando que essa pessoa nova contribua imediatamente, mas sem esperar que seja um especialista em manipulação de dados, comunicação empresarial e desenvolvimento de software ao mesmo tempo? Para isso, é preciso examinar a descrição e tentar entender o que o empregador está de fato buscando. Suponha que esteja olhando listas de adoção de gatos e que descrevem o gato Melão Cantaloupe como “ele gosta de perguntar sobre seu dia”. Seria necessário entender o que essa descrição realmente significa, se ele vai ficar miando por atenção, o que pode ser ruim para sua casa.

Nas descrições das vagas, algumas frases comuns para prestar atenção incluem “muito trabalho, muita diversão”, o que significa que terá de trabalhar longas horas e participar de eventos informais da empresa (como ir a bares) e “proativo e independente”, o que quer dizer que não terá muito suporte. Sabendo ler as entrelinhas, você garante que está se candidatando às vagas certas.

A primeira coisa a ter em mente é que as descrições das vagas são geralmente listas de desejo com alguma flexibilidade. Se você atender a 60% dos requisitos (talvez falte um ano do tempo de experiência exigido ou não tenha trabalhado com um recurso tecnológico da empresa), mas, sendo adequado ao seu perfil, ainda deve se candidatar à vaga. Não se preocupe muito com as competências desejáveis. Além disso, a exigência de anos de experiência profissional é apenas uma variável para as competências necessárias. Se você programou na universidade, essa experiência pode contar. Dito isso, candidatar-se a uma vaga para cientista de dados sênior que requeira cinco anos de experiência de trabalho como cientista de dados, proficiência em Spark e Hadoop e experiência na implantação de modelos de machine learning na produção provavelmente não seja o melhor uso de seu tempo se for um aspirante a cientista de dados proveniente do setor de marketing. A empresa está buscando um nível diferente de experiência e qualificações.

**REQUISITOS DE FORMAÇÃO** Muitas vagas de cientista de dados enumeram uma formação de “disciplina quantitativa” (campos como estatística, engenharia, ciência da computação ou economia) como requisito. Se você não tiver uma dessas formações, poderá se candidatar a essas

vagas? Em geral, sim. Discutimos esse tópico em mais profundidade no Capítulo 6, mas, se cursou disciplinas nessas áreas (incluindo um bootcamp ou na modalidade online), dá para enfatizar essa formação. Se você seguiu o conselho do Capítulo 4 de montar um portfólio e escrever posts de blog, é possível mostrar esses projetos aos empregadores evidenciando que consegue trabalhar com isso.

Uma complicação nas vagas de cientista de dados é que palavras diferentes podem significar a mesma coisa. Machine learning e estatística têm esse conceito. Uma empresa requer experiência em regressão ou classificação e outra requer experiência em aprendizado supervisionado, mas, em geral, esses termos são equivalentes. Também se aplica aos testes A/B, experimento online e ensaios controlados e randomizados. Se não estiver familiarizado com um termo, faça uma pesquisa. É possível que você se dê conta de que fez o mesmo trabalho com um nome diferente! Se não tiver trabalhado com uma tecnologia específica referenciada na vaga, veja se fez algo similar. Se a lista citar Amazon Web Services (AWS), por exemplo, e tiver trabalhado com o Microsoft Azure ou o Google Cloud, você tem a habilidade para trabalhar com serviços de computação em nuvem.

Outra vantagem de saber como decodificar uma descrição de vaga é a capacidade de detectar alertas (Seção 5.1.2). Nenhuma empresa dirá que é ruim trabalhar nela. Quanto mais cedo reconhecer uma provável situação de trabalho ruim, melhor. É bom começar a procurar sinais de alerta na descrição da vaga.

### **5.1.2 Como detectar sinais de alertas**

Encontrar um emprego é uma via de mão dupla. Nesse processo, pode sentir que as empresas têm todo o poder e que você precisa provar que é merecedor. Mas você – sim, você – também pode ser seletivo. Acabar parando em um local de trabalho tóxico ou com um trabalho entediante é uma situação muito difícil. Embora nem sempre seja possível saber se esse será o caso, vendo a descrição da vaga é possível prestar atenção em alguns sinais de alerta:

- *Sem descrição* – o primeiro sinal de alerta é não ter uma descrição da empresa ou da própria vaga – apenas uma lista de requisitos. Tais empresas esqueceram que contratar é um processo que envolve dois lados, e elas não estão pensando em você. Ou podem estar entrando na

moda da ciência de dados e querem apenas ter cientistas de dados sem fazer os ajustes necessários para que possam ser produtivos.

- *Requisitos amplos e extensos* – um segundo sinal de alerta é a descrição do unicórnio já mencionada (Seção 5.1.1). Mesmo que esse exemplo seja extremo, deve ter cuidado com descrições de vagas que tenham dois ou três dos tipos de trabalho (ciência da decisão, análise de dados e machine learning) como responsabilidades principais. Embora seja normal esperar uma competência básica em cada função, nenhuma pessoa poderá preencher todas essas funções no nível de especialista. Mesmo que alguém pudesse, não teria tempo para fazer tudo isso.
- *Desencontros* – por fim, procure desencontros entre os requisitos e a descrição da vaga. O empregador está pedindo experiência em deep learning, mas as funções do cargo incluem fazer dashboards, comunicar-se com stakeholders e conduzir experimentos? Em caso afirmativo, a empresa pode apenas querer alguém que possa utilizar a ferramenta do momento ou que seja um cientista de dados de “prestígio” com um doutorado em inteligência artificial, quando, na verdade, nem usa esse conhecimento especializado.

### 5.1.3 Definição das suas expectativas

Embora você deva ter algum padrão de exigência para uma vaga em potencial, não dá para exigir perfeição. Por vezes, os aspirantes a cientistas de dados veem seu caminho dividido desta forma: “Etapas 1-98: aprender Python, R, deep learning, estatísticas bayesianas, computação em nuvem, testes A/B, D3. Etapa 99: encontrar um emprego em ciência de dados. Etapa 100: lucrar”. Este exemplo é um exagero, mas parte do entusiasmo com a ciência de dados é a idealização de como é trabalhar na área. Afinal de contas, o cientista de dados é “o melhor emprego nos EUA” (<http://mng.bz/pyA2>), com um salário alto e uma satisfação elevada no trabalho. Você pode imaginar passar todos os dias nos problemas mais interessantes na área com os colegas mais inteligentes. Os dados de que necessita serão sempre acessíveis e limpos, e quaisquer problemas que enfrentar serão imediatamente resolvidos por uma equipe de engenheiros. Seu trabalho será exatamente como descrito, e nunca terá que trabalhar nas



partes da ciência de dados que menos lhe interessam.

Infelizmente, esse cenário é uma fantasia. Assim como esperamos que a Parte I deste livro o tenha convencido de que não é necessário saber tudo antes de entrar na área, as empresas também não serão unicórnios perfeitos. Há uma razão pela qual este livro não termina quando você consegue um emprego de ciência de dados. Embora tornar-se um cientista de dados seja uma grande realização e deva se orgulhar, a ciência de dados é uma área na qual estará sempre aprendendo. Os modelos falharão, a política no local de trabalho apagará o trabalho que vinha fazendo desde o último mês ou você passará semanas trabalhando com engenheiros e gerentes de produtos para coletar os dados de que necessita.

É especialmente fácil idealizar as empresas que são bem conhecidas, no geral ou na ciência de dados. Talvez você tenha ido a uma palestra e um dos funcionários da empresa falou algo genial. Talvez tenha seguido o blog dessa pessoa por meses e saiba que ela tem conhecimento de ponta na área. Pode ser que tenha lido um artigo que diz que a empresa tem cabines de dormir, refeições gourmet e vários escritórios que aceitam cachorros. Mas seja lá o que o tenha atraído, provavelmente também interessou a outros aspirantes a cientista de dados. A maioria dessas empresas recebe centenas de candidatos para uma vaga aberta e pode exigir mais do que o necessário para preenchê-la. Em todo caso, a vaga pode ser em uma divisão totalmente diferente e até pouco interessante.

Mesmo com expectativas realistas, é provável que você não tenha como primeiro emprego na ciência de dados algo dos seus sonhos. É mais fácil fazer a transição dentro da sua área ou trazer a ciência de dados para o seu trabalho do momento. Mesmo que esteja buscando sair da sua área, pode ser necessário começar indo para um cargo no qual possa aproveitar suas outras competências. Não significa que você não deve ter certos requisitos e preferências, mas é bom ter alguma flexibilidade. É muito normal mudar de emprego na área da tecnologia mesmo após um ou dois anos, ou seja, não está se candidatando para trabalhar nos próximos 15 anos. Mas você não pode saber exatamente o que quer antes mesmo de estar na área, e aprenderá até mesmo com os trabalhos ruins, então, não se estresse demais.

### 5.1.4 Participação de meetups

Embora os sites de vaga sejam uma forma comum de encontrar ofertas de trabalho, normalmente não são os locais mais eficazes para se candidatar. Como discutimos no Capítulo 6, enviar seu currículo online com frequência tem uma taxa de retorno muito baixa. De acordo com uma pesquisa do Kaggle em 2017 (<https://www.kaggle.com/surveys/2017>), as duas maneiras mais comuns de que pessoas já empregadas como cientistas de dados procuram e conseguem empregos é com recrutadores, amigos, familiares e colegas. Uma ótima maneira de formar uma rede de contatos é participar de meetups (uma mistura de encontro com palestra).

Meetups via de regra são realizados de modo presencial em alguma noite da semana. Normalmente, há um palestrante, um painel ou uma série de palestrantes que apresentam um tópico relevante para o evento. Os meetups, em geral, são gratuitos ou cobram uma pequena taxa, que, por vezes, serve para a comida. Alguns meetups podem ter apenas 20 pessoas; outros podem preencher um salão com 300. Alguns se reúnem todos os meses; outros, apenas algumas vezes por ano. Alguns encorajam os participantes a permanecerem no local depois da palestra ou para se encontrarem em um bar próximo; outros se concentram na própria palestra. Alguns têm focos muito específicos, como o processamento de linguagem natural avançada em Python; outros podem oferecer uma introdução a séries de tempo em um mês e modelos avançados de deep learning no próximo. Vale a pena participar de alguns meetups para ver de qual mais gosta. O tópico é importante, mas é bom encontrar um lugar onde se sinta bem-vindo e aprecie falar com os demais participantes. Quase todos os encontros têm contas em <https://www.meetup.com>, então, dá para pesquisar ciência de dados, machine learning, Python, R ou análise de dados para encontrar encontros relevantes em sua área.

Muitos meetups de ciência de dados têm um tempo reservado no início para que as pessoas anunciem se estão contratando. Converse com elas; o recrutamento faz parte do trabalho delas e, mesmo que as vagas abertas não sejam adequadas para você, podem dar bons conselhos ou sugerir outras possibilidades.

Você também pode encontrar outro participante que trabalhe na empresa

ou no setor em que esteja interessado. É possível perguntar se eles têm tempo para uma entrevista informativa para que possa aprender mais sobre a área. Uma entrevista informativa não é (ou, em vez disso, não deveria ser) uma forma passiva-agressiva de pedir indicação; em vez disso, é uma ótima forma de conferir uma empresa e obter conselhos de alguém que trabalha lá. Apesar de falarmos no Capítulo 6 sobre as vantagens de se ter uma indicação para uma vaga, não recomendamos pedir isso a pessoas que você acabou de conhecer. É pedir demais a alguém que eles não conhecem, e ninguém gosta de se sentir usado. Se falarem sobre uma vaga na empresa deles e disserem que podem indicá-lo, é um grande bônus, mas você ganhará muito com a realização de entrevistas informativas.

Participar de meetups é ótimo também por outras razões. Uma é encontrar pessoas que pensam como você e que estão na sua mesma região. Se tiver se mudado para uma nova cidade ou acabou de se formar na faculdade, pode se sentir como um estranho na cidade. Ir aos meetups é uma grande oportunidade de desenvolver sua carreira e construir um círculo social. Você pode aproveitar os meetups para montar uma rede de contatos que poderá ajudá-lo com qualquer coisa, desde questões específicas de ciência de dados a recomendações de busca de emprego ou orientação geral. Além disso, embora alguns meetups publiquem online gravações das palestras, outros não o fazem; por isso, comparecer é a única forma de ouvir essa palestra.

Infelizmente, os meetups podem ter algumas desvantagens. Pode ser intimidante participar de um meetup com poucas pessoas, todas com experiência e/ou que conheçam umas às outras. A síndrome de impostor pode surgir, mas você deve combatê-la, pois há poucos espaços mais acolhedores do que um bom meetup. Por fim, embora os meetups ofereçam uma ótima oportunidade de conhecer o cenário local da ciência de dados, eles podem ser isolados ou não ter muita diversidade, dependendo de como os organizadores são acolhedores e de como um meetup particular está conectado a uma comunidade diversa.

### **5.1.5 Utilização de redes sociais**

Se não residir em uma grande cidade ou perto de uma, pode não haver

meetups de ciência de dados perto de você. Nesse caso, o Twitter e o LinkedIn são excelentes locais para começar a montar sua rede de contatos. Ao seguir alguns cientistas de dados bem conhecidos, encontrará com frequência mais pessoas para seguir à medida que vê quem é retuíto ou quem é muito mencionado. Também pode começar a fazer um nome para si mesmo.

Gostamos de utilizar o Twitter de algumas maneiras:

- *Compartilhar trabalho* – quando escrever um post no blog, é bom que as pessoas o leiam! É totalmente normal se promover, fazendo uma ligação ao seu trabalho com uma breve descrição.
- *Compartilhar o trabalho de outras pessoas* – você leu algo muito bom? Encontrou um pacote que lhe poupou horas de frustração? Viu um slide numa palestra que foi particularmente útil? Se sim, ajude outras pessoas a serem iluminadas também. No Capítulo 6, discutimos uma das melhores formas de entrar em contato com alguém: mencione como você se beneficiou do trabalho dela. Marcar o criador em seu post é uma boa maneira de chegar no radar deles de forma positiva. Se compartilhar uma palestra, verifique se a conferência ou meetup tem uma hashtag; usar essa hashtag no seu tweet é uma ótima maneira de dar mais visibilidade.
- *Pedir ajuda* – há algum problema que tenha ficado travado no fato de que você (e o Google) não foi capaz de resolver? É provável que outra pessoa tenha enfrentado a mesma questão. Dependendo do tipo de problema, pode haver fóruns ou sites específicos onde dá para fazer perguntas ou pode fazer uma chamada geral usando uma hashtag relevante.
- *Dicas de compartilhamento* – nem tudo serve para o post de um blog, mas se tiver uma dica rápida, compartilhe-a. O tópico pode parecer algo que “todos” sabem, porém lembre-se: as pessoas que estão apenas começando não sabem tudo. Mesmo as pessoas que têm usado uma determinada linguagem há anos podem não saber sobre uma nova maneira de fazer alguma coisa.

Se sua busca de emprego não for sigilosa, também pode postar nas redes

sociais que está buscando e perguntar se alguém tem alguma indicação. Mesmo que ainda não tenha uma boa rede de contatos em ciência de dados, pode ter amigos, ex-colegas e outros contatos que possam saber de vagas nas empresas deles. Normalmente, essa abordagem funciona melhor nas plataformas de redes sociais onde as pessoas que estão ligadas à ciência de dados tendem a se reunir, como o LinkedIn ou o Twitter, mas até mesmo redes sociais, como o Facebook, podem ter contatos com oportunidades.

No início da carreira é comum sentir que você ainda não tem uma rede de contatos; as redes parecem ser mantidas por pessoas que já têm empregos na ciência de dados! A solução é fazer contatos não só quando estiver buscando um emprego, mas também muito antes disso. Quanto mais sair de sua zona de conforto e conversar com as pessoas ao seu redor em conferências, meetups, instituições acadêmicas, churrascos e afins, mais preparado estará na próxima vez em que buscar um emprego.

## **Como manter o pipeline cheio**

Um erro comum ao buscar emprego é depositar suas esperanças em uma única oportunidade e não continuar se candidatando e fazendo entrevistas em outros lugares. O que acontece se essa oportunidade não der certo? Não é necessário recomeçar do zero o processo de busca de emprego. É bom ter várias oportunidades em cada etapa: candidaturas enviadas, triagens de recursos humanos, estudos de caso e entrevistas presenciais. Não considere o processo terminado até que tenha aceitado uma oferta por escrito.

Ter várias oportunidades também ajuda a lidar com a rejeição. A rejeição é quase inevitável ao buscar um emprego, sendo difícil não levar para o lado pessoal ou como indicação do seu valor. Em alguns casos, você pode sequer ser notificado de que foi rejeitado; simplesmente nunca mais ouviu falar da empresa. Mas muitas razões pelas quais pode não ter conseguido o emprego estão fora de seu controle. A empresa pode ter fechado a vaga sem contratar ninguém, decidiu escolher um candidato interno ou aceitou alguém quando você ainda estava no início do processo. A rejeição dói, sobremaneira de uma empresa com a qual você estava realmente animado. Dê um tempo para processar seus sentimentos. Mas ter outras opções pode ajudá-lo a manter-se motivado e a fazer progresso.

Por fim, ter várias opções em potencial facilita lidar com a rejeição. Talvez tenha feito a triagem do RH e um estudo de caso para, então, descobrir que não há engenheiros de dados; a equipe de ciência de dados é formada apenas por algumas pessoas, mesmo que a empresa seja grande ou aquilo que a empresa está procurando é muito diferente daquilo que anunciou. Embora não deva esperar pela vaga perfeita de ciência de dados (isso não existe), você provavelmente tem alguns requisitos não negociáveis, e é muito mais fácil manter-se com esses requisitos se entender que eles podem ser atendidos em outras vagas.

## 5.2 Como decidir a quais vagas se candidatar

A essa altura, você deve ter uma lista de dezenas de vagas pelas quais está interessado e para as quais poderia ser um bom candidato. Você envia currículo a todas as vagas imediatamente?

Bem, algumas pessoas o fazem para dezenas ou mesmo centenas de vagas. Elas estão tentando jogar com as probabilidades, pensando que se houver uma chance de 10% de receber um retorno de uma delas, então, ao candidatar-se ao maior número possível, elas receberão mais retornos. Mas elas estão operando sob uma falácia: se você tem uma quantidade finita de energia e tempo, dividi-la em 100 candidaturas em vez de 10 torna cada uma delas mais fraca. Falaremos no Capítulo 6 sobre como adaptar seu currículo a cada cargo, mas só é possível se você for seletivo sobre a qual vaga se candidatar. É praticamente impossível fazer isso para 50 empresas.

**R E PYTHON** Você deve se candidatar a uma vaga se a empresa pedir Python e souber R ou vice-versa? Embora saber uma linguagem facilite aprender a outra, você já aprenderá muito no seu primeiro emprego como cientista de dados: trabalhar com stakeholders, política interna, estatística, conjuntos de dados e assim por diante. Mesmo se conseguisse o emprego, aprender uma linguagem nova no meio de tudo isso pode ser difícil. Assim, geralmente recomendamos que se candidate apenas a vagas que utilizem sua linguagem principal. Se saber uma dessas linguagens é apenas desejável, mas não um requisito, é bom ser cauteloso; essa descrição pode significar que não programaria. Por fim, algumas vagas pedem as duas linguagens. Você também deveria ser um pouco cauteloso aqui; normalmente, esse requisito significa que as pessoas usam qualquer linguagem, não que todos saibam ambas, o que pode tornar a colaboração difícil. Esse tipo de trabalho pode funcionar, mas pergunte nas entrevistas qual linguagem é mais utilizada. Se você for uma de apenas duas pessoas usando Python em uma equipe de 20, será difícil melhorar suas habilidades de programação.

Talvez seja melhor voltar ao que aprendeu nos dois primeiros capítulos

sobre os tipos de empresas de ciência de dados e sobre o trabalho de ciência de dados. É bom testar todas as partes diferentes da ciência de dados, ajustando um sistema de recomendação durante um mês e criando um modelo de vida vitalício no mês seguinte? Se sim, é provável que queira trabalhar para uma empresa que começou recentemente a fazer ciência de dados, já que empresas mais maduras terão funções especializadas. Por outro lado, grandes empresas de tecnologia também têm legiões de engenheiros de dados; então, obter dados de rotina é rápido e fácil.

Alguns desses fatos serão óbvios sobre a empresa; uma startup de 10 pessoas, por exemplo, não terá um sistema maduro de ciência de dados. Mas como saber mais?

Primeiro, verifique se a empresa tem um blog sobre ciência de dados. Em geral, somente as empresas de tecnologia possuem esse tipo de blog, mas ler esses posts não contribui muito para aprender sobre o que os cientistas de dados fazem de fato. Seus pensamentos positivos sobre posts específicos no blog de uma empresa também são ótimos para incluir na sua carta de apresentação para essa empresa (abordada no Capítulo 6). Se nunca ouviu falar de uma empresa, confira o site dela. Quando você sabe o que a empresa faz e como ela ganha dinheiro, pode começar a fazer suposições sobre que tipo de trabalho de ciência de dados ela necessita. Por fim, se estiver realmente interessado em uma empresa, veja se algum de seus cientistas de dados tem um blog no qual falam sobre o trabalho deles ou se deram palestras sobre isso.

Ao ler sobre uma empresa, lembre-se de pensar no que é geralmente importante para você. Poder trabalhar de casa é importante? E quanto ao número de dias de férias? Se quiser ir a alguma conferência, a empresa oferece reembolso de algumas despesas e dias livres para que você possa participar? Também ler o que a empresa diz sobre si mesma pode mostrar seus valores. Ela fala sobre jogos, cerveja no escritório e jantares? É provável que essa empresa esteja cheia de funcionários jovens. Ou enfatiza em horas de trabalho flexíveis ou férias familiares? Essa empresa mais provavelmente é amiga dos pais. No Capítulo 8, discutiremos como negociar muito mais do que o salário, mas, nesta etapa, você pode ao menos ver se a empresa anuncia benefícios que se alinham com suas prioridades.



Agora que você tem uma lista gerenciável de empregos em potencial, é hora de se candidatar! No Capítulo 6, explicamos como criar um bom currículo e uma carta de apresentação, incluindo como adaptá-los para cada vaga.

## **5.3 Entrevista com Jesse Mostipak, promotora na Kaggle**

Jesse Mostipak tem formação em biologia molecular e trabalhou como professora de escola pública antes de se apaixonar pela ciência de dados sem fins lucrativos. No momento desta entrevista, ela era diretora de ciência de dados na Teaching Trust. Você pode encontrar o que ela escreve sobre ciência de dados sem fins lucrativos, conselhos sobre aprender R e outros tópicos no site dela: <https://www.jessemaegan.com>.

### **Quais são suas recomendações para iniciar uma busca de emprego?**

Pense o quão dependente você está do título de cientista de dados. Se decidir não se preocupar em como pode ser chamado e se concentrar no trabalho que está fazendo, terá muito mais flexibilidade na busca de emprego. Algumas palavras-chave que não são de cientista de dados a serem pesquisadas são *análise*, *analista* e *dados*. Embora tenha mais vagas para filtrar, você pode encontrar um título como “pesquisa e avaliação”, em que está qualificado para essa vaga, mas nunca saberia se estivesse apenas pesquisado por cientista de dados.

Ao ler as vagas, concentre-se no que quer fazer como cientista de dados. Para mim, não sinto muito prazer em calcular o retorno sobre o investimento de cliques de site. Eu me perguntei: “Com quais causas eu me importo? Quais empresas estão alinhadas com isso?” Importava-me muito com as Girls Scouts (escoteiras) e, coincidentemente, estavam buscando um analista; então, consegui entrar e fazer isso. A mesma coisa aconteceu com a Teaching Trust quando eu quis ir mais para a área de educação.

### **Como montar uma rede de contatos?**

Quando estava fazendo a transição para a ciência de dados, fiz várias coisas que deram errado durante muito tempo. Eu era a pessoa que repostava no Twitter cada artigo de ciência de dados que via, fazendo 20 posts em um dia que não geravam nenhum engajamento. Você deve pensar sobre quem quer encontrar, por que, que valor você traz a esse relacionamento e o que é autêntico para você. Pense em criar uma marca para si mesmo, não necessariamente de maneira estrita, mas garantindo que a forma como você se mostra online e nas redes sociais seja autêntica. Para mim, percebi que não podia ser uma cientista de dados perfeita e estar nas redes sociais até saber de tudo sobre ciência de dados, porque esse dia nunca chegaria. Em vez disso, decidi falar sobre coisas que estava aprendendo e ser transparente sobre o processo. Foi assim que montei minha rede.

### **O que fazer quando se sentir inseguro para se candidatar a vagas na ciência de dados?**

Se estiver desenvolvendo competências, pode fazer algumas análises em Python ou R e terá os conceitos básicos sob controle. Concentre-se em como pode se sentir confortável com riscos e falhas. Você tem que falhar muito como um cientista de dados. Se estiver preocupado com o risco e a falha no processo de se candidatar a uma vaga, o que acontecerá quando assumir um risco em um modelo e ele não funcionar? Você precisa abraçar a ideia de ambiguidade e iteração. Candidate-se e tente; você será rejeitado, mas é normal! Sou rejeitada com certa regularidade. É apenas parte da experiência.

### **O que diria a alguém que pensa: “Não atendo à lista completa das qualificações necessárias das vagas”?**

Alguns estudos sugerem que determinados grupos de pessoas sentem que precisam estar 100% qualificados, enquanto outras dizem: “Cumpro 25%? Vou me candidatar!”. Atenha-se a essa confiança dos qualificados em 25% e tente. Mas também é possível ficar confuso com a linguagem da descrição da vaga. Por exemplo, digamos que há uma competência listada, como 10 anos trabalhando com bancos de dados SQL. Talvez pense: “Não tenho, mas tenho sete anos trabalhando com o Microsoft Access”. Eu diria que é

uma competência transferível. Cabe ao candidato dizer a si mesmo: “Posso não ter exatamente essa competência, mas tenho uma competência muito semelhante. Preciso dar uma olhada no SQL, ver como minha competência é transferível e dizer a essa empresa as coisas incríveis que fiz com o Microsoft Access e que eles devem me contratar porque sei que posso fazer isso com o SQL”.

## **Qual é seu último conselho para aspirantes a cientistas de dados?**

Você precisa desenvolver suas competências de comunicação e se deixar levar. Deve ser capaz de se comunicar com todos os setores da empresa de uma forma que respeite a experiência das pessoas com quem estiver falando e também mostrar que você existe para facilitar a vida delas.

Com *se deixar levar* quero dizer algo como: “Não é assim que eu lidaria com esse problema ou este projeto, mas consigo enxergar seu ponto de vista. Vamos tentar dessa forma e talvez eu possa modificá-lo dessa maneira”. Você também precisa ser flexível, porque as pessoas estão recém descobrindo a ciência de dados. As empresas querem cientistas de dados, mas não sabem muito bem o que fazer com eles. Você está à mercê da empresa; se as necessidades dela mudarem, é preciso evoluir e adaptar-se para melhor atender a essas necessidades.

Por fim, saiba que a descrição da vaga pode mudar. Você poderia dizer: “Não é isso que pensei que faria, mas como posso fazer funcionar para mim?”. Não pode dizer: “Sou o melhor em redes neurais, mas não estou trabalhando com redes neurais, por isso, obviamente, este trabalho é um lixo”. É preciso saber que todo trabalho que você assume mudará e evoluirá conforme as necessidades da empresa.

## **Resumo**

- Pesquise termos gerais como *dados* em sites de vagas e concentre-se nas descrições, não nos títulos.
- Não se preocupe em cumprir com 100% das qualificações listadas.
- Lembre-se de que o processo de busca de emprego é uma via de mão

dupla. Preste atenção em sinais de alerta e pense sobre que tipo de ciência de dados quer fazer.

## CAPÍTULO 6

# Como se candidatar: currículo e carta de apresentação

Este capítulo abrange:

- Como fazer um currículo convincente e uma carta de apresentação
- Como adaptar sua candidatura a cada vaga

Você tem uma lista de vagas abertas pelas quais está interessado; agora é hora de informar aos empregadores que você existe! Quase todas as vagas requerem o envio do currículo: uma lista de suas competências e experiência. A maioria das vagas também pede uma carta de apresentação: uma carta de uma página que descreve por que você deve ser considerado para a vaga. Pode ser fácil falar de experiências prévias e dizer que está interessado na empresa, mas, nesta situação, esforçar-se mais pode ser o fator decisivo para conseguir uma entrevista.

Neste capítulo, começamos garantindo que seu currículo e sua carta de apresentação são os mais eficazes possíveis, tratando de práticas recomendadas e erros comuns a serem evitados. Depois, mostramos como pegar esse currículo “modelo” e a carta de apresentação e refiná-los para cada vaga. Por fim, mostramos como uma rede de contatos pode ajudar a colocar sua candidatura feita com cuidado nas mãos de um gerente de recrutamento, em vez de entrar em uma grande pilha de currículos.

**OBSERVAÇÃO** O único objetivo de um currículo é convencer uma pessoa que dá uma olhada rápida em seu currículo que vale a pena entrevistá-lo.

O tema principal deste capítulo é que você precisa convencer uma pessoa rapidamente de que está qualificado para a vaga. Os recrutadores das empresas recebem com frequência centenas de currículos para cada vaga aberta em ciência de dados. Além disso, como a ciência de dados abrange

muitos tipos distintos de vagas, o leque de competências dos candidatos às posições será enorme. Esse ato reforça a noção de que seus documentos têm que dizer: “Ei, você que está lendo isto aqui, pode parar de olhar essa pilha enorme, porque já encontrou a pessoa com as competências que está procurando”. No entanto, ser capaz de mostrar que está qualificado não é uma tarefa fácil.

Embora uma rede de contatos e personalizar sua candidatura levem tempo, isso rende resultados muito melhores do que gastar apenas uma hora fazendo uma carta e um currículo básicos e, depois, candidatar-se para dezenas de vagas com apenas um clique. É mais provável conseguir uma entrevista, pois terá correspondido aos requisitos da empresa. Quando chegar à entrevista (tópico do Capítulo 7), poderá dar uma ótima resposta à pergunta comum: “Por que você está interessado nesta função?”.

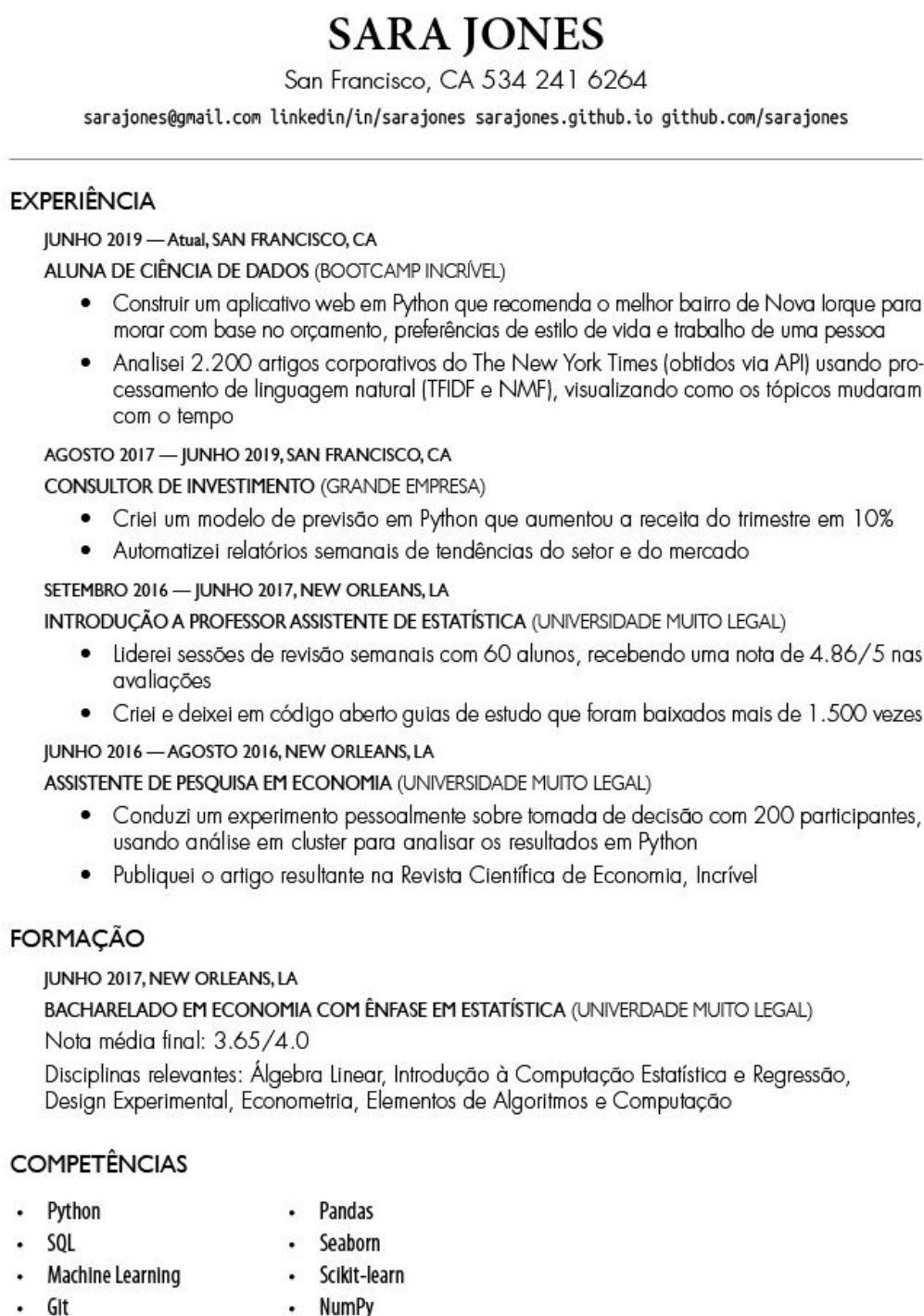
## **6.1 Currículo: o básico**

O objetivo do seu currículo não é conseguir o emprego, mas, sim, uma entrevista. Os recrutadores do processo de entrevista ficam em apuros se trazem candidatos que claramente não cumprem as qualificações para a vaga, e eles são elogiados quando as pessoas se encaixam bem nas qualificações. Seu currículo precisa mostrar ao recrutador que você atende aos requisitos para a vaga de modo que ele ficará confortável em seguir com o processo de seleção.

Esse objetivo é muito diferente de criar um catálogo de cada experiência que você teve, que é, infelizmente, um objetivo que muitas pessoas inexperientes têm ao fazer um currículo. Embora queira evitar lacunas no currículo, deixando de fora empregos recentes, pode gastar menos tempo naqueles que não estão relacionados à ciência de dados. Mesmo que tenha muita experiência em ciências de dados, continue concentrando-se em destacar os empregos mais relevantes. Se tiver um currículo de muitas páginas, a maioria dos recrutadores não terá tempo para ler tudo, nem poderá saber quais partes ler. Ninguém dirá: “Teríamos contratado você, mas não mencionou seu emprego de atendente quando estava no ensino médio, então não deu”.

Depois, haverá oportunidade no processo da entrevista de falar de todos seus empregos, formação e de ciência de dados em maior profundidade. Por agora, concentre-se no que é mais relevante para atender às qualificações do cargo para o qual está se candidatando. Durante o restante do processo, você se concentrará em suas grandes qualidades que o ajudarão a se destacar dos demais candidatos, mas, para a primeira etapa, é bom concentrar-se na adequação às expectativas do gerente da contratação ou do recrutador.

Com isso em mente, vamos tratar da estrutura básica de um currículo, como criar um bom conteúdo dentro dessa estrutura e como pensar nas muitas regras de currículo que existem. Muito desse conteúdo pode se aplicar a qualquer vaga técnica do setor, mas iremos nos concentrar tanto quanto possível no que é único em relação à ciência de dados. Também elaboramos um exemplo de currículo que serve como guia (Figura 6.1).



*Figura 6.1 – Exemplo de currículo para um aspirante a cientista de dados.*

### 6.1.1 Estrutura

Nesta seção, falaremos de cada parte de um currículo exemplar, entrando em mais detalhes sobre o que incluir.



## Seção: informações de contato

SARA JONES

San Francisco, CA 534 241 6264

[sarajones@gmail.com](mailto:sarajones@gmail.com) [linkedin/in/sarajones](https://www.linkedin.com/in/sarajones) [sarajones.github.io](https://sarajones.github.io) [github.com/sarajones](https://github.com/sarajones)

É necessário incluir suas informações de contato para que o recrutador possa contatá-lo. Inclua seu nome e sobrenome, número de telefone e email, no mínimo. Além disso, é possível colocar links de locais onde possam encontrar mais informações sobre você, incluindo perfis de redes sociais, como LinkedIn, bases de códigos online, como o GitHub, e sites e blogs pessoais. Para pensar no que adicionar, faça esta pergunta: “Se clicassem neste link, pensariam melhor de mim?”. Um link para seu portfólio de projetos do Capítulo 4, por exemplo, é algo fantástico para ser incluído, mas um link para um perfil do GitHub em branco, exceto por uma cópia (clone) de um projeto de tutorial, não é. Se tiver algum trabalho de ciência de dados publicamente disponível, tente descobrir uma maneira de mostrá-lo aqui.

Em geral, é bom incluir a cidade e o estado onde mora, o que permite que o recrutador saiba se você está próximo e se pode se deslocar para o trabalho ou se precisa se mudar do lugar se assumir a vaga. Algumas empresas hesitam em fazer realocação em novas contratações devido à despesa; por isso, se não morar nas proximidades e não quiser mostrar isso de cara, pode deixar sua localização de fora.

Se seu nome legal não corresponde ao nome que normalmente utiliza, coloque seu nome mais usado. Mais adiante no processo, precisará informar à empresa seu nome legal para verificações de antecedentes, mas não é necessário usá-lo ao se candidatar.

Por fim, não use um endereço de email potencialmente ofensivo ([odeio\\_Python@gmail.com](mailto:odeio_Python@gmail.com), por exemplo) ou algum que possa expirar (como o email da universidade).

## Seção: experiência

JUNHO 2019 — Atual, SAN FRANCISCO, CA

ALUNA DE CIÊNCIA DE DADOS (BOOTCAMP INCRÍVEL)

- Construir um aplicativo web em Python que recomenda o melhor bairro de Nova Iorque para morar com base no orçamento, preferências de estilo de vida e trabalho de uma pessoa
- Analisei 2.200 artigos corporativos do The New York Times (obtidos via API) usando processamento de linguagem natural (TFIDF e NMF), visualizando como os tópicos mudaram com o tempo

Nesta seção você mostra que está qualificado para a vaga por causa de empregos, estágios ou bootcamps que teve ou fez. Se seus empregos prévios estão relacionados com a ciência de dados, como engenharia de software, é ótimo: use uma boa parte do seu currículo com eles. Se um emprego não estiver relacionado à ciência de dados, como ser professor de história da arte, você ainda deve listar os empregos, mas não passe muito tempo neles. Para cada cargo ocupado, liste o nome da empresa, o mês e o ano de início e término, o título do cargo e pelo menos um ponto importante (dois ou três para empregos mais relevantes) descrevendo o que fez. Se for recém-formado, pode incluir estágios e trabalhos de pesquisa realizados na faculdade.

Esta seção deve ser a maior em seu currículo, podendo ocupar metade do espaço disponível. Muitas vezes também é a mais importante, porque é o primeiro lugar que os recrutadores irão ler para verificar se você tem experiência em ciência de dados que pode estar relacionada com a vaga aberta. Devido à importância de preencher corretamente essa parte, iremos nos aprofundar sobre como criar o melhor conteúdo para ela na Seção 6.1.2.

## Seção: formação

FORMAÇÃO

JUNHO 2017, NEW ORLEANS, LA

BACHARELADO EM ECONOMIA COM ÊNFASE EM ESTATÍSTICA (UNIVERSIDADE MUITO LEGAL)

Nota média final: 3.65/4.0

Disciplinas relevantes: Álgebra Linear, Introdução à Computação Estatística e Regressão, Design Experimental, Econometria, Elementos de Algoritmos e Computação

Nesta seção, liste suas experiências educacionais, para mostrar que tem um conjunto de competências que seriam úteis para o trabalho de ciência de dados. Se você foi para a universidade depois do ensino médio, mesmo que não tenha se formado, liste as escolas/universidades, datas (mesmo formato da sua experiência profissional) e área de estudo. Se essa vaga for a primeira fora da universidade, e sua média de notas do curso for elevada, pode inclui-la; caso contrário, deixe de fora. Se for recém-formado e tiver feito estatística, matemática ou ciência da computação ou qualquer outra

disciplina relacionada (como métodos de pesquisa em ciências sociais ou engenharia), pode listar essas disciplinas.

Os recrutadores estarão muito interessados em ver se você tem uma área de estudo que seja relevante para as ciências de dados, como formação em ciências de dados, estatística, ciência da computação ou matemática. Eles também estarão interessados em ver o nível da sua formação. Como muitos tópicos de ciência de dados não são abordados até chegar no nível de pós-graduação, ter um diploma de pós-graduação ajuda. Em geral, recrutadores geralmente não se preocupam em qual universidade você estudou, a menos que seja extremamente famosa ou de prestígio, e, mesmo assim, essa credencial não importa se já estiver formado há alguns anos. No entanto, é bom que os recrutadores vejam quaisquer bootcamps, certificados ou programas online, porque mostram que você deu continuidade à sua formação.

Embora a seção de formação de seu currículo possa dar informações valiosas ao recrutador, você não pode melhorar a seção fazendo nada além do que estudar outro curso ou receber algum certificado, abordado no Capítulo 3.

## Seção: competências

### COMPETÊNCIAS

- Python
- SQL
- Machine Learning
- Git
- Pandas
- Seaborn
- Scikit-learn
- NumPy

Esta seção é onde pode listar explicitamente todas as competências relevantes que você tem a contribuir em um ambiente de ciência de dados. O ideal é que o recrutador veja esta seção e diga: “Sim, muito bom!”, porque terá listadas as competências que são relevantes para o trabalho. Em currículos de ciência de dados, há dois tipos de competências a serem listadas nesta seção:

- *Competências de programação/banco de dados* – essas competências podem ser linguagens de programação como Python e SQL, frameworks e ambientes como .NET ou JVM, ferramentas como Tableau e Excel ou ecossistemas como Azure e Amazon Web Services (AWS).

- *Métodos de ciência de dados* – o segundo tipo trata de métodos de ciência de dados, como regressões e redes neurais. Como é viável listar muitos métodos possíveis, tente focar em alguns importantes que mostram que você tem os fundamentos e algumas competências especiais. Algo como “regressões, métodos de clustering, redes neurais, análise de pesquisas”, por exemplo, mostra que tem os princípios e a profundidade.

Tente não listar mais de sete ou oito competências para evitar sobrecarregar as pessoas. Não liste competências que não tenham chance de serem relevantes para a vaga (como uma linguagem de programação acadêmica obscura do tempo na universidade).

Liste apenas as competências com as quais se sentiria confortável em usar no trabalho, e não uma linguagem que não usa há cinco anos e não quer usar novamente. Se algo está no seu currículo, é justo que o recrutador possa perguntar. Se as vagas em ciência de dados que você viu solicitarem certas competências – e que você tem – lembre-se de incluí-las! Essa informação é exatamente o que os recrutadores procuram.

Recomendamos não usar classificações por estrelas, números ou outros métodos para tentar indicar o quão bom você é em cada competência. Por um lado, as classificações não significam nada: qualquer pessoa pode se classificar em 5/5. Se der pontuações perfeitas, os recrutadores podem pensar que não é honesto ou bom em autoavaliação; se atribuir pontuações mais baixas, podem duvidar de suas competências. Além disso, não fica claro como você considera cada nível. Será que 5/5 significa que acha que é um dos melhores do mundo, que sabe fazer uma tarefa avançada ou que é melhor em certas competências do que seus colegas de trabalho? Se um recrutador quiser uma autoavaliação do seu nível de competências, ele lhe perguntará em uma entrevista.

Não enumere competências sociais, como pensamento crítico e competências interpessoais; embora sejam cruciais para ser um cientista de dados bem-sucedido, colocá-las no currículo não tem sentido, pois qualquer pessoa pode fazê-lo. Se quiser destacar suas competências nessas áreas, fale sobre como as usou em momentos específicos na seção de experiência do currículo. Além disso, não precisa listar as competências básicas que

qualquer pessoa que se candidate possa ter, como trabalhar com o pacote Microsoft Office.

### **Seção: projetos de ciência de dados (opcional)**

Se tiver feito projetos de ciência de dados fora do trabalho, poderá criar uma seção para eles. Esta seção é ótima para candidatos com menos experiência de trabalho, mas que tenham feito projetos paralelos ou em uma universidade ou bootcamp. Basicamente, está dizendo ao recrutador: “Embora não tenha muita experiência de trabalho relevante, não importa, porque realizei um processo completo de ciência de dados”.

Para cada projeto você precisará de um título, da descrição do que fez e de como o fez e dos resultados. Na verdade, os projetos de ciência de dados devem ser como se fossem trabalhos em estrutura e conteúdo; então, tudo na Seção 6.1.2 sobre geração de conteúdo aplica-se a eles também. Idealmente, terá um link para um post de blog ou pelo menos para um repositório do GitHub que tenha um arquivo README informativo. A ciência de dados é um campo técnico em que é fácil mostrar o trabalho feito, e esta seção é um ótimo lugar para isso. Se já tiver experiência de trabalho relevante, pode pular esta seção, mas ainda deve falar sobre projetos nas entrevistas.

### **Seção: publicações (opcional)**

Se tiver publicado artigos relacionados à ciência de dados em um programa de mestrado ou doutorado, deve incluí-los. Se tiver publicado artigos em outras áreas, mesmo quantitativos, como física ou biologia computacional, pode incluí-los, mas apenas brevemente. Como não estão diretamente relacionados à ciência de dados, a pessoa que os lê não entende muito das publicações, exceto que trabalhou bastante para ser publicado. Você pode listar o trabalho relevante que realizou durante sua pesquisa na seção de experiência, como: “Criei um algoritmo para analisar milhões de sequências de RNA por minuto”. Mas uma publicação em um periódico do qual o recrutador nunca ouviu falar, mesmo que seja de prestígio na sua área, não ajuda muito.

### **Outras seções**

É possível adicionar outras seções, como láureas acadêmicas e prêmios, se tiver ganho competições do Kaggle ou recebido uma bolsa de estudos, mas não são necessárias. Você não precisa incluir referências; a parte de falar com as referências virá mais tarde no processo, e pode compartilhar essas informações se chegar lá. Declarações objetivas geralmente não são necessárias e são redundantes, dadas as demais informações em seu currículo. A frase “cientista de dados experiente em Python procurando um cargo para desenvolver capacidades de modelagem e teste A/B”, por exemplo, não anima um recrutador.

## **Montando o quebra-cabeças**

Geralmente, colocam-se as informações de contato na parte superior, seguidas da seção mais importante. Se estiver na universidade ou for recém-formado, provavelmente precisará colocar sua formação primeiro; se não tiver trabalho ou formação relevante, coloque seus projetos de ciência de dados primeiro; caso contrário, coloque sua experiência de trabalho. Nas seções de trabalho e formação, liste suas experiências na ordem cronológica reversa, da mais recente a menos recente.

Já vimos muitos formatos eficazes de currículos de ciências de dados. Neste campo, você tem um pouco de liberdade no design; não há exatamente um formato padrão. Apesar dessa liberdade, é bom focar em tornar seu currículo fácil de digitalizá-lo rapidamente. Como os recrutadores gastam tão pouco tempo olhando seu currículo, não é bom que usem esse tempo tentando descobrir como encontrar seu trabalho mais recente. Não faça o design ser uma distração para o conteúdo; considere como os outros o verão. Algumas boas práticas incluem:

- Cabeçalhos simples para as seções, a fim de facilitar encontrá-las
- Mais espaço em branco para facilitar a leitura do conteúdo
- Colocar em negrito palavras importantes, como os cargos em cada empresa

Se ideias como espaços em branco e cabeçalhos são muito para você, atenha-se a um modelo de currículo encontrado online ou consulte um especialista em design.

Em geral, é bom limitar seu currículo a uma única página. Essa prática

tem duas finalidades: como seu currículo será lido rapidamente, isso serve para garantir que o recrutador invista esse tempo na informação que você pensa que é a mais valiosa, e mostra que pode se comunicar de maneira concisa e compreender quais partes de sua experiência são as mais importantes de compartilhar. Se alguém enviar um currículo de 17 páginas (o que já vimos), é um grande indício de que a pessoa não faz ideia do que a torna uma boa candidata e que ela acha que as pessoas teriam tempo de ler tudo isso.

Por fim, certifique-se de ser consistente em seu currículo. Se abreviar os meses na seção de formação, abrevie também na seção de experiência de trabalho. Embora possa usar fontes e tamanhos diferentes para cabeçalhos e texto de corpo, não mude o formato de cada linha do texto da descrição. Use verbos no passado para cargos anteriores e no presente para o cargo atual. Essas coisas mostram que presta atenção a detalhes e (novamente) ajuda os leitores a processarem seu conteúdo rapidamente, já que não ficarão distraídos com as mudanças de fonte ou estilo. É improvável que uma única inconsistência faça você perder uma entrevista, mas, às vezes, os detalhes fazem toda a diferença.

**REVISÃO GRAMATICAL** É essencial revisar seu currículo! Alguns erros ortográficos ou gramaticais podem colocar seu currículo na lixeira (metafórica). Por que tão rigoroso? Quando os recrutadores estão lendo centenas de currículos, dois tipos se destacam: aqueles que são claramente excepcionais (raros) e aqueles que são fáceis de eliminar. Este último tipo precisa de algumas regras básicas, e além de currículos que claramente não satisfazem os requisitos, os currículos com erros ortográficos são uma razão fácil para eliminar um candidato. As vagas de ciência de dados requerem atenção a detalhes e verificação do trabalho; se não conseguir fazê-lo ao dar o melhor de si ao se candidatar, o que esse fato sugere acerca do seu trabalho? Além de usar o recurso de verificação ortográfica do seu processador de texto, peça para pelo menos uma outra pessoa ler seu texto com atenção.

### **6.1.2 Como aprofundar a seção de experiência: geração de conteúdo**

Esperamos que pegar as datas e títulos do seu histórico de trabalho e formação seja bem fácil. Mas como criar o conteúdo para descrever sua experiência de trabalho (ou projetos de ciência de dados)?

O erro comum que as pessoas cometem nos currículos é a criação de apenas uma lista de tarefas, como “relatórios gerados para executivos que

usam SQL e Tableau” ou “ministrei a disciplina de cálculo a três turmas de 30 alunos”. Há dois problemas com essa abordagem: ela indica apenas pelo que era responsável, não o que realizou ou como o fez, e pode ser enquadrada de uma forma que não seja relevante para a ciência de dados. Para os dois exemplos anteriores, poderia descrever o mesmo trabalho como “automatizei a geração de relatórios de previsões de vendas para executivos usando Tableau e SQL, economizando quatro horas de trabalho por semana” ou “ensinei cálculo a 90 alunos, obtendo uma média de 9,5/10 nas avaliações dos alunos, com 85% dos alunos recebendo uma nota 4 ou 5 no exame AP de Cálculo BC”.

Tanto quanto possível, é bom explicar sua experiência em termos de competências que são transferíveis para a ciência de dados. Mesmo que não tenha trabalhado em ciência ou análise de dados, houve algum dado com o qual trabalhou? Recrutadores estão dispostos a considerar a experiência fora das funções da ciência de dados como relevante, mas você tem que explicar o motivo pelo qual devem. Se algum trabalho estiver relacionado a pegar dados e entendê-los, deve se esforçar em criar uma história concisa sobre o que fez. Você analisou 100 GB de dados de estrelas para um doutorado em astrofísica? Gerenciou 30 arquivos de Excel para planejar os funcionários de uma padaria? Muitas atividades envolvem o uso de dados para entender um problema.

Você usou ferramentas como Google Analytics, Excel ou Survey Monkey? Mesmo que essas ferramentas não sejam aquelas que a vaga está pedindo, trabalhar com dados de qualquer tipo é relevante. Que competências de comunicação usou? Explicou conceitos técnicos ou da área, talvez em palestras de pesquisa de doutorado ou em empresas? Se as competências transferíveis forem difíceis, não se preocupe; o restante dos conselhos sobre como escrever melhores descrições ainda ajudará. Mas se ainda não o fez, deve pensar em como sua formação ou seus projetos paralelos podem demonstrar suas competências em ciências de dados, especialmente se sua experiência profissional não puder.

Para os cargos menos relevantes ocupados alguns anos atrás, está tudo bem em ter apenas uma breve descrição. De um modo geral, não deixe um emprego de fora do seu currículo, de maneira que fique um intervalo de



mais de alguns meses. Se está trabalhando há algum tempo e teve muitos empregos, está tudo bem em listar apenas os três ou quatro mais recentes.

Talvez ache que esse processo é muito mais fácil no seu trabalho de agora do que aquele que você tinha cinco anos atrás. Uma boa prática é manter uma lista do que realizou e dos principais projetos em que trabalhou. Quando você está no dia a dia do trabalho, fazendo progresso pouco a pouco, é possível esquecer-se de quão impressionante foi todo o trabalho realizado ao olhar para trás. As pessoas sabem que seu currículo não é uma lista exaustiva, então, elas não pensam que “demorou 15 meses para montar um sistema automatizado para rastrear e classificar oportunidades de vendas que pouparam à equipe de vendas mais de 20 horas de trabalho manual por semana”. Elas pensam: “Uau, precisamos de um sistema como este!”.

Em geral, as descrições caem em duas categorias. A primeira é de grandes conquistas, como “criei um dashboard para monitorar todos os experimentos em execução e realizar cálculos de potência”. A segunda é uma média ou total, como: “Implementei e analisei mais de 60 experimentos, resultando em mais de 30 milhões de dólares em receitas adicionais”.

Em qualquer um dos casos, cada descrição deve começar com um verbo e (idealmente) ser quantificável. Em vez de dizer: “Fiz apresentações para clientes”, escreva: “Criei mais de 20 apresentações para executivos da Fortune 500”. É ainda melhor se puder quantificar o impacto que teve. Escrever “executei 20 testes A/B em campanhas de email, resultando em um aumento de 35% na taxa de cliques e em um aumento de 5% nas vendas atribuídas em geral” tem muito mais impacto do que “executei 20 testes A/B em campanhas de email”.

## **6.2 Carta de apresentação: o básico**

Embora o propósito de um currículo seja oferecer aos recrutadores fatos relevantes sobre sua experiência de trabalho e formação, o objetivo da carta de apresentação é ajudá-los a entender quem você é como pessoa. Sua carta de apresentação é onde pode explicar como encontrou a empresa e destacar o motivo pelo qual é um bom candidato. Se seu currículo não mostrar um

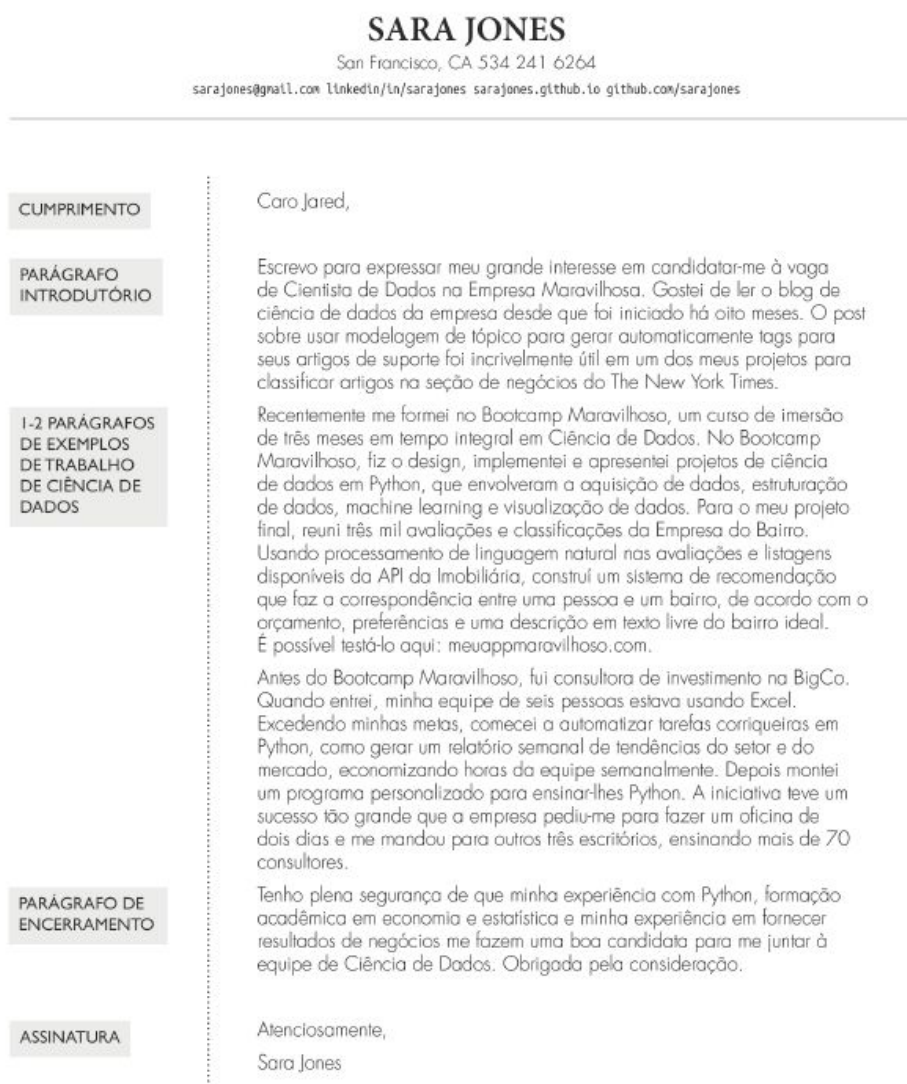
caminho linear, a carta de apresentação pode explicar como as peças se encaixam para que você seja um forte candidato à vaga. Até mesmo apenas mostrar que sabe o que a empresa faz, que visitou o site dela ou que usou seu produto (se estiver disponível para pessoas físicas) pode ajudar. Sua carta de apresentação é a melhor ferramenta para auxiliar os recrutadores a entenderem coisas que não cabem nas descrições das experiências.

Ao contrário de um currículo, uma carta de apresentação pode ser opcional. Se uma empresa tiver um lugar para incluir uma, envie. Algumas empresas eliminam os candidatos que não escreveram uma carta. Não é raro que as empresas deem um tema específico para escrever, como sua técnica de aprendizado supervisionado favorita. Esse pedido é normalmente feito para verificar se as pessoas leram e seguiram o pedido da descrição da vaga, em vez de enviar uma carta de apresentação genérica para todo mundo. É bom mostrar à empresa que você consegue seguir instruções.

Sabendo que uma carta de apresentação serve para ajudar a empresa a entender melhor quem você é, um erro comum que vemos em nessas cartas é concentrar-se no que a empresa pode fazer por você. Não diga: “Este seria um grande passo para a minha carreira”. O trabalho de um recrutador não é ajudar o maior número possível de carreiras, mas, sim, contratar pessoas que possam auxiliar a empresa. Mostre-lhes como pode fazer isso. Mesmo que fosse seu primeiro emprego de ciência de dados, que experiência relevante você tem? Que resultados (mesmo que não estejam relacionados com a ciência de dados) você pode partilhar para que fique claro que trabalha arduamente e alcança objetivos? Não se subestime. Tente ter uma visão geral sobre como tornar-se atraente para a empresa.

Assim como seu currículo, sua carta de apresentação deve ser curta; pouco mais de meia página a uma página costuma ser a regra. Foque nos seus pontos fortes. Se a descrição da vaga lista quatro competências, e você se sai muito bem em duas, fale sobre essas duas! Não fique sentindo que tem de pedir desculpas pelas competências que faltam.

A Figura 6.2 mostra um exemplo de carta de apresentação.



*Figura 6.2 – Exemplo de carta de apresentação com destaques que mostram diferentes componentes.*

## 6.2.1 Estrutura

As cartas de apresentação têm regras menos bem definidas do que os currículos. Dito isso, eis uma boa estrutura geral que pode seguir:

- **Saudação** – Se possível, descubra quem é o recrutador da vaga. O primeiro lugar para pesquisar é a própria descrição da vaga; o nome dessa pessoa pode estar ali. Mesmo que tenha só um endereço de email, provavelmente consegue descobrir quem é o recrutador com uma pesquisa online. Caso contrário, confira o LinkedIn e o site da empresa para ver se em um deles identifica o líder da equipe. Mesmo que acabe

pensando grande demais – talvez o vice-presidente do departamento que será o gerente do seu gerente –, você ainda demonstrou que fez alguma investigação. Por fim, se não conseguir encontrar um nome, enderece sua carta ao *Prezado gerente do departamento X* (como em *“Prezado gerente de análise de dados”*). Não use “Prezado senhor ou senhora”; essa frase é arcaica e genérica.

- *Parágrafo introdutório* – apresente-se, nomeie a vaga para a qual está interessado e explique brevemente por que está animado com a empresa e a função. Se a empresa tiver um blog de ciência de dados ou se algum dos cientistas de dados tiver palestrado ou feito posts de blog sobre o trabalho, esse parágrafo é um ótimo lugar para mencionar que assistiu ou leu. Conecte o que aprendeu nessas apresentações com o motivo pelo qual você está animado com a vaga ou a empresa.
- *Um a dois parágrafos de exemplos de trabalho em ciência de dados* – conecte suas realizações anteriores com essa função. Adicione detalhes a algo discutido em seu currículo e aprofunde-se em uma função ou projeto paralelo, dando exemplos específicos. Siga o princípio “mostre, não diga”; em vez de escrever que você é um “solucionador de problemas, organizado e orientado a detalhes”, forneça um exemplo de ser esse tipo de pessoa no trabalho.
- *Parágrafo de encerramento* – agradeça ao recrutador pelo tempo e consideração. Resuma suas qualificações explicando por que você seria um bom candidato à vaga.
- *Assinatura* – “Atenciosamente” e “Obrigado(a) por sua consideração” são bons fechamentos. Evite escrever algo muito casual ou muito simpático, como “Até mais” ou “Abraço”.

## 6.3 Adaptação

As duas seções anteriores trataram de regras gerais para a redação de uma carta de apresentação e de um currículo eficazes. Mas a melhor maneira de se diferenciar de outros candidatos é adaptar esses documentos à posição para a qual está se candidatando.

A primeira pessoa que lê seu currículo de ciência de dados provavelmente

não será o gerente do cargo; pode nem mesmo ser um humano! Em empresas maiores, os sistemas de acompanhamento de candidatos automaticamente avaliam currículos buscando palavras-chave, deixando em evidência aqueles que contêm essas palavras. Tal sistema pode não reconhecer “modelagem linear” como algo que satisfaça a exigência de experiência em “regressão”. Talvez um leitor humano também não consiga; uma pessoa de recursos humanos pode ter recebido apenas a descrição da vaga e instruções para encontrar candidatos promissores. É bom não arriscar que um recrutador não entenda que seu projeto usando “vizinhos mais próximos ao k” (k-nearest neighbors) significa que tem experiência em análise de cluster ou que *NLP* é o acrônimo de *natural language processing* – *processamento de linguagem natural*. É bom que alguém possa olhar para seu currículo e a descrição da vaga facilmente, encontrando correspondências exatas para os requisitos de sua experiência. Embora não seja recomendável sobrecarregar seu currículo com jargão técnico, é bom usar palavras-chave (como *R* ou *Python*) algumas vezes para ajudar o currículo passar por essas primeiras avaliações.

Recomendamos ter um currículo e uma carta de apresentação “modelos”, que possam ser adaptados em vez de começar do zero toda vez. Essa abordagem é especialmente útil se estiver se candidatando para tipos diferentes de vagas. Se alguns trabalhos enfatizarem machine learning e outras análises exploratórias, será muito mais fácil se colocar termos-chave relacionados na descrição das suas experiências. Seu currículo e a carta de apresentação modelos podem ter mais de uma página, mas certifique-se de que os enviados tenham no máximo uma página.

Adaptar sua candidatura à vaga não significa que precisa ter uma descrição ou uma competência para cada requisito. Como discutimos no Capítulo 5, as descrições da vaga geralmente são listas de desejo; tente pensar quais são as competências principais para o trabalho. Por vezes, de forma útil, as empresas dividem as competências e a experiência em “requisitos” e “desejável”, mas mesmo se não fizeram isso, é possível saber qual é qual lendo a descrição das responsabilidades. Embora as empresas gostem de contratar alguém que tenha todos os requisitos e mais um pouco, a maioria não se prende a isso.

Uma exceção são as grandes empresas de tecnologia e startups conhecidas e em rápido crescimento. Essas empresas recebem muitos candidatos e procuram razões para rejeitar as pessoas. Elas estão muito preocupadas com falsos-positivos, o que significa contratar alguém que é ruim ou apenas mediano. Elas não se importam com falsos-negativos – não contratar alguém que é ótimo – porque elas têm muita gente no pipeline. Para essas empresas, você geralmente tem que atender a 90% dos requisitos ou até mesmo 100%.

## 6.4 Indicações

Todos os sites de vagas e de empresas têm um lugar para se candidatar, às vezes com um clique de um botão se você salvou seu currículo no site. Infelizmente, como é tão fácil se candidatar dessa forma, seu currículo acaba com frequência em uma pilha de centenas ou mesmo milhares de candidaturas frias similares. É por isso que recomendamos não se candidatar dessa forma até que tenha esgotado outras opções. Ler a descrição de vagas é uma ótima maneira de ter uma ideia de que tipo de empregos estão disponíveis, mas a melhor maneira de colocar seu pé na porta é ter alguém para abri-la para você.

É bom tentar usar a porta escondida da maioria das empresas: indicações. Uma *indicação* significa um funcionário da empresa indicar alguém para uma vaga, geralmente enviando a candidatura e informações dessa pessoa por um sistema especial. Muitas empresas oferecem bônus de indicação, pagando a funcionários se eles indicarem alguém que aceite uma oferta de trabalho. As empresas gostam de pessoas que são indicadas porque já vêm pré-avaliadas: alguém que já trabalha na empresa e (presumivelmente) está indo bem, pensa que essa pessoa seria um bom candidato. Mesmo que alguém não o indique formalmente, ser capaz de escrever na sua carta de apresentação “conversei sobre essa vaga com o [Funcionário X]” e que essa pessoa possa dizer ao gerente para olhar seu currículo são uma grande vantagem.

Como encontrar pessoas que possam te indicar? Comece olhando no LinkedIn para ver se conhece alguém que trabalha em uma empresa na qual

está interessado. Mesmo que não tenha falado com essa pessoa há algum tempo, tudo bem em entrar em contato enviando uma mensagem educada. Depois, procure pessoas que trabalharam anteriormente na mesma empresa ou que estudaram na mesma universidade que você. É mais provável receber retorno de uma mensagem se mencionar algo que vocês tenham em comum. Por fim, procure pessoas que são contatos de segundo grau para ver quem vocês têm em comum. Se tiver uma boa relação com qualquer uma das suas conexões mútuas, entre em contato com essa pessoa para ver se ela estaria disposta a apresentá-lo.

Se estiver buscando um cientista de dados, tome algum tempo para aprender sobre o que fazem. Essas pessoas têm um blog, uma conta no Twitter ou um repo no GitHub onde compartilharam o trabalho delas? Mark Meloon, cientista de dados da ServiceNow, escreveu no seu blog: “Como subir degraus das relações para conseguir um emprego em ciência de dados (texto em inglês)” (<http://mng.bz/O95o>) que as mensagens mais eficazes são aquelas que combinam um elogio ao conteúdo publicado com um pedido para fazer mais algumas perguntas. Dessa forma, também evitará perguntar coisas sobre as quais já conversaram publicamente e pode se concentrar em receber conselhos que não encontraria em outros lugares.

Lembre-se de que não são apenas as pessoas na ciência de dados que podem ajudá-lo. Embora outros cientistas de dados estejam mais bem posicionados para lhe dizer como é trabalhar na empresa deles, pessoas em qualquer vaga podem indicá-lo. Se alguém que você conhece trabalha em uma empresa para a qual quer se candidatar, entre em contato com ele! No mínimo, podem dar uma ideia sobre a cultura da empresa.

## **Como escrever uma mensagem eficaz**

No post de blog “Você tem tempo para uma conversa rápida?” (texto em inglês) (<http://mng.bz/YeaK>), Trey Causey, gerente sênior de ciência de dados na Indeed.com, descreve algumas sugestões para, de maneira eficaz, entrar em contato com alguém que você não conhece para falar sobre seu projeto, busca de emprego ou escolha de carreira. Ao seguir essas diretrizes, é muito mais provável que receba uma resposta, tenha uma reunião produtiva e construa uma boa base para uma relação contínua:

- Faça uma lista sobre o que quer conversar e inclua-a no email.
- Sugira alguns horários (incluindo um tempo de término 30 minutos depois) e um local próximo ao trabalho da pessoa.
- Pague pelo almoço ou café da pessoa.
- Chegue cedo.
- Tenha perguntas e objetivos específicos para a conversa com base na lista enviada. Não peça apenas “qualquer dica que possa me dar”.
- Controle o tempo e avise quando o tempo passar; se a pessoa quiser continuar conversando, ela dirá.
- Agradeça à pessoa e busque mais informações sobre o que conversaram.

A seguir, como Trey junta tudo numa mensagem de exemplo:

“Olá, Trey, li o post do blog sobre entrevistas de ciência de dados e gostaria de convidá-lo para um café no Storyville em Pike Place esta semana a fim de fazer algumas perguntas sobre seu post.

Estou fazendo entrevistas, e a parte sobre codificação em quadro branco (whiteboard) me interessou muito. Gostaria de saber o que você acha de como melhorar as perguntas e respostas sobre este assunto, bem como compartilhar algumas das minhas experiências com esses tipos de perguntas.



Você teria 30 minutos em algum momento – digamos, terça ou quarta-feira da próxima semana? Agradeço o post!”

## **6.5 Entrevista com Kristen Kehrer, instrutora de ciência de dados e criadora de curso**

Kristen Kehrer é instrutora de ciência de dados na Universidade da Califórnia – Berkeley Extension, membro do corpo docente do Emeritus Institute of Management e fundadora da Data Moves Me, LLC. A Data Moves Me ajuda as equipes de ciência de dados a comunicarem os resultados do modelo de machine learning aos stakeholders para que a empresa possa tomar decisões com confiança. Ela tem mestrado em estatística aplicada e é coautora do livro *Mothers of Data Science* (autopublicado).

### **Em uma estimativa, quantas vezes acha que editou seu currículo?**

Nossa, um milhão! Venho de uma família de trabalhadores normais, meu pai era bombeiro e minha mãe, dona-de-casa, por isso nunca me ensinaram a escrever um grande currículo para a indústria. Mas deu tudo certo pedindo ajuda a outras pessoas quando estava saindo da universidade. Também sempre tive o hábito de manter registros de qualquer novo projeto em que eu trabalho ou qualquer coisa interessante que poderia adicionar ao meu currículo. Eu não sou daquelas pessoas que passaria dois anos sem atualizar o currículo. Mais recentemente, minha empresa anterior pagou por um coach de carreira quando me demitiram. Tive de aprender tudo sobre as recomendações de como fazer currículos e como me posicionar de forma eficaz para conseguir um bom trabalho.

Recomendo muito que as pessoas atualizem o currículo com frequência. Em especial se você trabalha no mesmo lugar há um tempo, é muito difícil pensar em todas as coisas relevantes que poderia adicionar ao currículo. Por exemplo, fui coautora de alguns anúncios premiados no setor de saúde. Não é relevante para todos os cargos aos quais me candidatei, mas, se me candidatasse a uma vaga na saúde, posso fazer referência a essa pesquisa.

Se eu não tivesse mantido registros, não me lembraria de quem eram meus coautores ou qual era o título do anúncio.

### **Quais são os erros comuns que vê as pessoas cometerem?**

Tantas coisas! Um é o currículo de quatro páginas em que ainda constam trabalhos da época do ensino médio. Outro é não perceber que os sistemas de acompanhamento de candidatos não analisam certas coisas muito bem. Se pessoas tiverem ícones ou gráficos no currículo, isso não será reconhecido nos sistemas automatizados mais antigos e pode acabar com você sendo automaticamente rejeitado. Também não gosto quando as pessoas colocam, digamos, três estrelas em Python, porque não há um contexto, e para qualquer competência que colocar duas estrelas está dizendo que não é bom nisso.

### **Você faz adaptações ao seu currículo conforme a vaga à qual está se candidatando?**

Não sou obsessiva a esse respeito. Mas quase todas as empresas de grande e médio portes usam agora um sistema de acompanhamento de candidatos, e é bom usar as palavras-chave relacionadas. Se tivesse visto coisas em uma descrição de vaga específica que combinava com o que eu tenho de experiência, mas que talvez tenha usado palavras um pouco diferentes, eu editaria algumas palavras para corresponder à terminologia utilizada na descrição da vaga.

### **Quais estratégias recomenda para descrever experiências em um currículo?**

Digo às pessoas para arrumarem o currículo conforme a vaga que elas querem, e não para o trabalho que já têm. Não precisa fazer uma lista de todas as coisas que já fez. Em vez disso, pense no que fez e que pode transferir para a ciência de dados. Por exemplo, se for professor de matemática, você tem explicado material técnico ou matemático a um público não técnico. Ou talvez tenha trabalhado em um projeto onde, mesmo que não estivesse em análise de dados, teve que trabalhar de forma multifuncional em várias equipes. Ou seja, é bom mostrar que é capaz de

resolver problemas, autogerenciar-se, comunicar-se bem e alcançar resultados. Por fim, é possível usar projetos paralelos para destacar suas competências técnicas e proatividade.

## **Qual é seu último conselho para aspirantes a cientistas de dados?**

Você precisa começar a se candidatar a vagas de ciência de dados. Muitas pessoas simplesmente continuam fazendo cursos online porque acham que precisam conhecer um milhão de coisas para se tornarem cientistas de dados, mas o fato é que você vai começar um trabalho e ainda terá muito para aprender. Mesmo com dez anos na área, ainda tenho muito a aprender. Ao se candidatar, receberá feedback do mercado. Se ninguém der um retorno, talvez você não esteja se posicionando bem ou talvez não tenha as competências necessárias. Junte feedbacks de algumas pessoas e escolha uma área para focar, como ser capaz de automatizar processos em Python. Trabalhe nisso, adicione ao currículo e candidate-se mais vezes. Você precisa se candidatar, obter retornos, repetir e seguir até conseguir um trabalho.

## **Resumo**

- Seu currículo não é uma lista exaustiva de tudo o que já fez. Ele é necessário para conseguir uma entrevista, não o emprego, então foque em correspondê-lo à descrição da vaga.
- Cartas de apresentação permitem mostrar por que você está interessado em uma empresa e como sua experiência o posiciona para fazer uma contribuição valiosa.
- Falar com pessoas que atualmente trabalham em uma empresa, em especial se forem cientistas de dados, é a melhor maneira de conseguir informações específicas sobre vagas abertas e a cultura da empresa.

## CAPÍTULO 7

# A entrevista: o que esperar e como lidar com ela

Este capítulo abrange:

- O que os entrevistadores estão buscando
- Tipos comuns de perguntas na entrevista
- Etiqueta apropriada ao se comunicar com uma empresa

Se parar para pensar sobre o processo de uma entrevista, dá para perceber o quão complicado é: de alguma forma, é preciso mostrar a pessoas que não o conhecem que você seria bom em uma função sobre a qual sabe apenas alguns parágrafos da descrição da vaga. Na entrevista, podem perguntar questões técnicas de todos os níveis sobre diferentes tecnologias – algumas das quais você talvez nunca tenha usado antes. Além disso, durante a entrevista, precisará aprender o suficiente sobre a empresa para poder decidir se *quer* trabalhar lá. É preciso fazer tudo isso em apenas algumas horas, agindo com profissionalismo e de maneira apropriada. É o suficiente para sofrer de ansiedade!

A boa notícia é que, com a preparação e a mentalidade certas, as entrevistas de ciências de dados podem deixar de induzir ataques de pânico e serem gerenciáveis, toleráveis e talvez até mesmo uma experiência agradável.

Neste capítulo, vamos explicar o que os entrevistadores buscam e como ajustar suas ideias para se alinharem com as suas necessidades. Discutimos questões técnicas e não técnicas, bem como um estudo de caso em ciência de dados. Por fim, vamos ver como se comportar e quais perguntas devem ser feitas aos entrevistadores. Com essa informação, você deve estar bem preparado para o que vem adiante.

## 7.1 O que as empresas querem?

Quando os funcionários de uma empresa entrevistam candidatos para uma vaga, eles estão buscando uma pessoa crucial:

*Alguém que consiga fazer o trabalho.*

É o único tipo de pessoa que buscam. As empresas não estão à procura da alguém que acerta mais perguntas da entrevista ou quem tem uma maior formação ou anos de experiência. Só querem alguém que consiga fazer o trabalho que precisa de ser feito e que ajude a equipe a avançar nos objetivos.

Mas o que é necessário para fazer o trabalho? Bem, algumas coisas:

- *Ter as competências necessárias* – as competências necessárias podem ser técnicas e não técnicas. Na parte técnica, é preciso entender as competências que falamos no Capítulo 1: alguma combinação de matemática e estatística, bem como bancos de dados e programação. Já na parte não técnica, precisa de perspicácia empresarial geral, além de competências como gestão de projetos, gestão de pessoas, design visual e outras que são relevantes para a função.
- *Ser alguém sensato de se trabalhar* – se você falar algo ofensivo, agir defensivamente ou ter falhas de caráter que dificultem a interação ou colaboração com outras pessoas, a empresa não irá querer contratá-lo. Significa que durante a entrevista (e sempre, na verdade), é bom ser agradável, compassivo e positivo. Não significa que seu entrevistador irá querer beber uma cerveja com você. Significa apenas que sua futura equipe precisa vê-lo como alguém com quem queira trabalhar.
- *Ser capaz de fazer as coisas* – não basta ter as competências necessárias para fazer o trabalho; é preciso ser capaz de utilizá-las! Você precisa conseguir encontrar soluções para problemas no trabalho e implementá-las. A ciência de dados tem muitas situações nas quais uma pessoa pode ficar travada, como entender dados complicados, pensar no problema, testar modelos diferentes e organizar um resultado. Uma pessoa que possa superar cada um desses desafios será muito melhor em fazer o trabalho do que alguém que fica esperando ajuda sem pedir. As pessoas que tentam fazer tudo perfeito também têm problemas: nunca

concluírem uma tarefa significa não poder usá-la.





Sabendo que essas são as três coisas que as empresas buscam nos candidatos, você está pronto para entrar no processo de entrevistas. À medida que avançamos pelo processo de como a entrevista funciona e as perguntas que muitas vezes surgem, iremos enquadrar nossa discussão em relação a essas três ideias.

### **7.1.1 O processo de entrevista**

Embora o processo exato de uma entrevista de emprego varie conforme a empresa, as entrevistas tendem a seguir um teste padrão básico. Esse padrão é pensado para maximizar a quantidade de informação que uma empresa aprende sobre o candidato ao minimizar a quantidade de tempo que as pessoas da empresa levam para conduzir a entrevista. Os entrevistadores normalmente estão ocupados, têm muitas entrevistas para fazer e querem permitir comparações justas entre os candidatos, por isso o processo é simplificado e consistente. Aqui está um esboço básico do que esperar em um processo de entrevista (Figura 7.1):

1. *Triagem inicial por telefone* – essa triagem normalmente se trata de uma entrevista por telefone de 30 minutos (ou, por vezes, de uma hora) com um recrutador técnico: alguém que tem muita experiência na triagem de candidatos e conhece os termos técnicos, mas não faz o trabalho técnico. Do ponto de vista da empresa, o objetivo dessa entrevista é verificar se você tem alguma chance de estar qualificado para o trabalho. O recrutador quer filtrar pessoas que claramente não seriam boas para a vaga, como aquelas que não são exatamente qualificadas (não têm as competências necessárias) ou que parecem rudes ou mesquinhas (e não trabalhariam bem com outras pessoas). Da parte técnica, o entrevistador confere se você tem as competências mínimas necessárias, não se é o melhor para eles. É muito mais provável que perguntem: “Já usou regressões lineares?” do que “Como calcular a estimativa de probabilidade máxima para uma distribuição gama?”. Após a primeira entrevista por telefone, a empresa pode fazer outra entrevista também por telefone com um entrevistador mais técnico. Se a triagem inicial por telefone for boa, dentro de algumas semanas, você

fará...

	1. Entrevista por telefone
	2. Entrevista presencial
	3. Estudo de caso
	4. Entrevista com liderança e proposta

*Figura 7.1 – As quatro etapas do processo de entrevista.*

2. *Uma entrevista na empresa* – essa entrevista muitas vezes leva de duas a seis horas e é a parte principal do processo de entrevista. Nessa visita, você conhecerá onde e com quem trabalharia e dá à empresa tempo para fazer perguntas que avaliem mais sua formação, suas competências e suas expectativas como cientista de dados. Você será entrevistado por várias pessoas na visita, cada uma fazendo perguntas sobre tópicos diferentes, alguns técnicos e outro não. O objetivo dessa entrevista é garantir (por meio de perguntas técnicas) que você tem as competências necessárias e (por meio de perguntas comportamentais e como se comporta) que é uma pessoa sensata com quem se trabalhar. Se a entrevista correr bem, é hora de...
3. *Um estudo de caso* – você receberá a descrição de um problema do mundo real e de dados relacionados a ele. Terá tempo na empresa ou em casa no fim de semana para analisar os dados, tentar resolver o problema e criar um relatório sobre ele. Em seguida, apresentará o relatório à equipe de contratação. Esse exercício mostra à equipe que você tem as competências necessárias (pela qualidade do seu relatório) e que pode fazer as tarefas (pelo que fez no relatório). Nem todas as empresas exigem essa etapa; às vezes, elas a substituem por uma apresentação sobre seu trabalho prévio. Se o relatório de estudo de caso correr bem, você terá...
4. *Uma entrevista final com a liderança* – essa entrevista é com o gerente sênior, diretor ou algum outro líder da equipe. O objetivo é que o líder o

aprove para a vaga e a equipe. Mesmo se a entrevista com a liderança acontecer primeiro, significa que a equipe de ciência de dados o considerou um bom candidato; então, é raro que essa entrevista rejeite a aprovação. Observe que essa entrevista com frequência ocorre logo após o estudo de caso, mas pode ocorrer no início ou no fim da primeira entrevista na empresa. Presumindo que tudo correrá bem, você receberá uma proposta em menos de duas semanas!

Quando o processo envolve todas essas etapas, em geral uma proposta é recebida entre três semanas e dois meses depois de ter enviado o currículo. Como pode ver, cada parte do processo de entrevista é pensada para um objetivo diferente na empresa.

Nas seções seguintes, vamos nos aprofundar em cada parte do processo, exemplificando como demonstrar suas competências e capacidades em cada cenário.

## **7.2 Etapa 1: a triagem inicial por telefone**

A primeira interação com a empresa provavelmente será por uma ligação telefônica de 30 minutos com um recrutador. É importante dar uma boa primeira impressão. Dependendo do tamanho da empresa, no entanto, a pessoa que ligou pode não estar relacionada à equipe de ciência de dados em que você trabalharia; então, *o objetivo é mostrar à empresa que poderia fazer esse trabalho, não necessariamente que seja a melhor pessoa para realizá-lo.*

Por quê? A pessoa com quem você vai falar nessa entrevista telefônica tem o trabalho de filtrar os candidatos não qualificados. Quando o recrutador fala com um candidato, ele está tentando avaliar se vale a pena que alguém da equipe de ciência de dados fale com o mesmo. Muitas vezes, as pessoas se candidatam a vagas para as quais não têm as competências (ou mentem dizendo que as têm), por isso o recrutador quer evitar que elas avancem no processo. Como candidato, seu objetivo é fazer com que o recrutador entenda que você está pelo menos minimamente qualificado para a vaga.

É provável que o recrutador faça estes tipos de perguntas:



- *Fale um pouco sobre você* – (Não é tecnicamente uma pergunta, mas é tratada como sendo.) O recrutador está pedindo uma visão geral de um ou dois minutos da sua história. Eles querem ouvir experiências que se relacionam com a vaga. Se estiver se candidatando a uma vaga de ciência da decisão, por exemplo, eles querem ouvir sobre sua formação e experiências ou projetos que analisou. É importante que sua resposta tenha de um a dois minutos. Se sua resposta for inferior a um minuto, pode parecer que não tem muito o que mostrar; se levar mais de dois minutos, pode parecer que não sabe resumir uma história.
- *Com que tipos de tecnologias está familiarizado?* – O recrutador está confirmando que você tem a experiência técnica para o trabalho. Além dessa pergunta específica, é de se esperar algumas sobre sua familiaridade com matemática e estatística, bancos de dados e programação, além de áreas empresariais (consulte o Capítulo 1). Também é bom mencionar todas as tecnologias que você sabe que estão relacionadas à vaga ou foram mencionadas na descrição da vaga. Se não tiver exatamente o que a vaga requer (como Python em vez de R), está tudo bem; basta ser aberto e honesto com relação a isso. Se acontecer de conhecer mais do que os recursos tecnológicos da empresa, talvez por ter conversado com pessoas que trabalham lá, tente formatar sua resposta em torno disso.
- *Por quais motivos está interessado nessa vaga?* – O recrutador quer entender o que o atraiu para a empresa em primeiro lugar. Uma resposta particularmente bem pensada mostra que você faz sua tarefa de casa e consegue fazer o trabalho. Responder: “Só cliquei em ‘candidatar-se’ em todas as vagas de ciência de dados no LinkedIn” não seria de bom senso. Não pense demais nesse tipo de pergunta; apenas demonstre que sabe o que a empresa faz e que tem um interesse genuíno na função. Na medida do possível, tente relacionar a função com suas formações e seus interesses.

Embora o recrutador faça perguntas sobre você e sua história, a ligação também é o momento para entender melhor a própria vaga. O recrutador deve passar pelo menos 10 minutos falando sobre a função e a equipe para a qual está sendo entrevistado. Nesse meio tempo, faça perguntas para ter

certeza de querer o trabalho e para que possa demonstrar um interesse sincero pela vaga. Essas perguntas podem incluir tópicos como viagens, cultura da empresa, como a equipe está mudando, as prioridades da equipe e por que a vaga foi aberta.

É possível que, durante a ligação, o recrutador tente entender suas expectativas salariais diretamente (“quanto espera receber de salário?”) ou indiretamente (“quanto você ganha atualmente?”). Pelo lado positivo, o recrutador está confirmando se a empresa pode pagar suas expectativas salariais, para não perder tempo entrevistando alguém que não aceitaria a proposta da empresa. Pelo lado negativo, o recrutador pode estar tentando limitar o salário a um valor abaixo do que a empresa poderia oferecer, porque você está dando informações sobre suas expectativas. Se possível, evite discutir o salário até que tenha avançado mais no processo. O Capítulo 8 abrange as negociações salariais.

Durante o telefonema, pergunte qual é o próximo passo do processo e qual é o cronograma. O recrutador deve dizer algo como: “Daqui uma semana teremos a resposta se podemos marcar uma entrevista”. “Quais são as próximas etapas?” é algo totalmente razoável de se perguntar. Não pergunte ao recrutador diretamente se passará para a próxima entrevista. Essa pergunta pode ser desconfortável para ele, e é provável que, de qualquer forma, não tenha autoridade para tomar a decisão.

Se a entrevista por telefone der certo, você será chamada para uma entrevista na empresa.

## **7.3 Etapa 2: entrevista na empresa**

Essa etapa é a principal do processo seletivo. A empresa o chamou para fazer uma entrevista e reservou várias horas. Você tirou um dia livre, vestiu roupas mais elegantes do que o normal e seguiu para a empresa.

### **O que vestir na entrevista**

Um aspecto das entrevistas muito discutido e debatido é o que vestir. Para as vagas em ciência de dados, é uma questão complicada pelo fato de os empregos poderem ser em diferentes setores corporativos, cada um com sua própria cultura e código de vestimenta. O que é completamente apropriado para uma entrevista pode ser totalmente inadequado para outra.

A melhor estratégia é perguntar ao recrutador na hora de marcar a entrevista o que vestir e qual é o código geral de vestimenta da empresa. O recrutador quer seu sucesso e não irá enganá-lo. Outra alternativa é falar com alguém que trabalha nessa empresa ou em uma similar. Se nada disso der certo, presuma que os setores burocráticos (finanças, defesa, saúde etc.) têm códigos de vestimentas mais formais, e as empresas jovens ou de tecnologia (startups, empresas de tecnologia em massa) têm códigos de vestimenta mais flexíveis. Evite extremos (sandálias, bermudas, vestidos de festa, chapéus); use algo confortável.

O objetivo da entrevista é ajudar a empresa a entender se você teria capacidade de fazer o trabalho para o qual seria contratado, e se você o faria bem. Dependendo da empresa e da vaga, entre três a dez pessoas devem ser chamadas para a entrevista, cada uma com pontos fortes e fracos. A empresa quer encontrar a pessoa que seria a melhor opção ou a primeira pessoa entrevistada que seria muito boa no trabalho.

Ao longo da entrevista, reforce a ideia de que pode fazer o trabalho bem. Esse conceito é diferente de ser o candidato mais inteligente, com mais anos de experiência ou que usou mais tipos de tecnologia. Em vez disso, é interessante ser o candidato que tem um equilíbrio saudável, de ser alguém sensato com quem se trabalhar, que tenha as competências suficientes para o trabalho e que pode fazer o que for solicitado.

O que acontecerá nessas várias horas de entrevista? Uma entrevista na empresa geralmente inclui o seguinte:

- *Uma visita ao local de trabalho e apresentação à equipe* – a empresa quer que você entenda como seria trabalhar nesse local e, potencialmente, impressioná-lo com bebidas e lanches gratuitos (se disponíveis). Essa parte leva menos de 15 minutos, mas é bom dar uma boa olhada para ver se ficaria feliz em trabalhar lá. As estações de

trabalho são tranquilas e fáceis de trabalhar? As pessoas parecem razoavelmente felizes e amigáveis? Os notebooks são antigos? Ao fazer essa visita, haverá conversas rápidas com alguém da empresa. Cuidado! Essas conversas fazem parte da entrevista; se soar desagradável ou rude, é possível perder a proposta de emprego. É bom parecer agradável, mas, o mais importante, ser autêntico (exceto se seu autêntico seja ser um idiota).

- *Uma ou mais entrevistas técnicas* – essa parte pode levar de 30 minutos a muitas horas, dependendo de quão rigorosa é a empresa. Questionarão sobre muitos tópicos e talvez precise trabalhar em um quadro branco ou computador. (Damos mais detalhes sobre essas perguntas mais adiante, na Seção 7.4.1 e no Apêndice.) O objetivo dessa entrevista não é para o entrevistador entender os tópicos mais profundos que você conhece ou se pode resolver os problemas mais complicados. O objetivo é ver se você tem as competências minimamente necessárias para fazer o trabalho, então seu objetivo é mostrar que sim.
- *Uma ou mais entrevistas comportamentais* – o objetivo das entrevistas comportamentais é ver como é sua relação com os outros e como consegue fazer as coisas. Farão muitas perguntas sobre suas experiências, incluindo como lidou com situações difíceis e como garantiu que os projetos se concretizassem. Poderão fazer perguntas de trabalho muito gerais, como: “Fale sobre uma situação na qual lidou com um colega difícil” até perguntas mais específicas de ciência de dados, como, por exemplo: “De que maneira lidar com um projeto de ciência de dados em que o modelo falha?”. Essas perguntas não terão necessariamente respostas certas ou erradas; elas estarão abertas à interpretação pelo entrevistador.

**DICA** Antes do dia da entrevista na empresa, é possível perguntar ao recrutador como será a entrevista. No mínimo, eles devem oferecer uma programação mostrando com quem conversará e quais tópicos serão abordados. Se perguntar, talvez deem alguns detalhes específicos sobre o que esperar das partes técnicas e comportamentais da entrevista. Essa informação irá ajudá-lo a chegar preparado.

Passar por uma entrevista na empresa tem um grande peso emocional. Você terá de mudar rapidamente entre pensar em tópicos técnicos para perguntas pessoais e sobre seus sonhos, tudo de forma profissional e amigável.

Dependendo do tamanho da empresa, uma ou várias pessoas podem entrevistá-lo, e é bom passar uma boa impressão para cada uma delas. Uma das melhores formas de controlar os nervos durante uma entrevista é lembrar que os entrevistadores também são pessoas e que eles querem que se saia bem, tanto quanto você. Eles são seus aliados, não seus adversários.

Nas seções a seguir, adentramos mais profundamente em diferentes partes da entrevista. Esta seção do capítulo trata de partes diferentes da entrevista na empresa, mas consulte o Apêndice para ver uma lista detalhada de exemplos de perguntas e respostas de entrevista, além de detalhes sobre como pensar a respeito das perguntas.

### **7.3.1 A entrevista técnica**

Para muitos cientistas de dados, uma entrevista técnica é a parte mais assustadora de todo o processo de entrevista. É fácil imaginar-se travado em frente a um quadro branco e sendo questionado sobre algo que não faz ideia de como responder, sabendo que não conseguirá o trabalho. (Só escrever essa frase já causou ansiedade às autoras!)

Para entender melhor sobre como lidar com a entrevista técnica, é necessário repensá-la. Se leu o Capítulo 4 deste livro e criou um portfólio de ciência de dados, você já passou na entrevista técnica. O ponto da entrevista é ver se tem as competências necessárias para ser um cientista de dados e, por lógica, você as tem porque fez ciência de dados! Se durante uma entrevista pensar não ser capaz de responder a uma pergunta complicada, é um sinal de que o entrevistador está fazendo um trabalho ruim, não você. Você tem as competências e a experiência necessárias. Essa parte da entrevista é pensada para que demonstre esses fatos. Se a entrevista não permitir isso, não é culpa sua.

Nesse processo, você tenta mostrar ao entrevistador que *tem as competências necessárias para o trabalho*. Mostrar que tem um conjunto de competências é uma atividade muito diferente de responder corretamente a cada pergunta. Uma pessoa pode dar exatamente as respostas que um entrevistador quer ouvir e ainda se sair mal na entrevista ou pode dar respostas incorretas e se sair bem. Considere duas respostas para uma pergunta de entrevista:

*Entrevistador: O que é a validação cruzada com dobra K (k-fold cross validation)?*

*Resposta A: Dividir aleatoriamente dados em grupos K de mesmo tamanho e utilizá-los como dados de teste para modelos K.*

*Resposta B: Selecionar aleatoriamente uma amostra de dados e utilizá-la como dados de teste para um modelo K vezes. Depois você pega a média dos modelos e a utiliza. Esse método serve para lidar com sobreajuste porque há um grupo de modelos com dados diferentes de treinamento. Eu usei esse método no projeto principal em meu portfólio para prever preços de casa.*

Tecnicamente, a resposta A está correta e a resposta B, não (é validação cruzada, mas não tecnicamente com dobra  $k$  porque os dados não foram divididos em grupos de mesmo tamanho). Dito isto, a resposta A não deu ao entrevistador muita informação, além de que o entrevistado sabia a definição; já a resposta B mostrou que o candidato sabia o termo, sabia por que foi usado e tinha experiência prática. O exemplo ilustra por que é tão importante, durante uma entrevista, pensar sobre como transmitir o fato de que tem as competências.

Especificamente, é possível fazer algumas coisas na entrevista técnica de ciência de dados para transmitir ao entrevistador que você tem estas competências:

- *Explicar seu pensamento* – na medida do possível, não dê apenas uma resposta; explique também por que razão sabe dessa resposta. Dar uma explicação mostra ao entrevistador como você pensa sobre o assunto e pode mostrar que está no caminho certo mesmo que não tenha dado a resposta correta. Uma palavra de cuidado: embora repetir a pergunta em voz alta possa ser útil (como: “Hum... uma regressão linear funcionaria aqui?”), alguns entrevistadores podem tomar esse comportamento como um sinal de que você não sabe muito. Pratique responder às perguntas diretamente. Também pratique como formular seu processo de raciocínio desde o início.
- *Faça referências às suas experiências* – falando sobre projetos ou trabalhos que já fez, você está mantendo a conversa em suas

competências práticas do mundo real. Essa abordagem pode tornar uma resposta ambígua mais plausível e concreta ou apenas fornecer um tópico alternativo de conversa se sua resposta estiver um pouco fora do tema. No entanto, use essa abordagem com moderação. Se passar tempo demais falando sobre seu passado em vez de falar sobre a pergunta, pode parecer que está evitando a questão.

- *Seja aberto e honesto se não souber a resposta* – é totalmente possível (e normal!) não saber a resposta de cada pergunta na entrevista. Tente se antecipar e explique o que você *de fato* sabe sobre a resposta. Por exemplo, se perguntaram: “O que é ‘semi join’?”, e não souber a resposta, pode dizer algo como: “Não ouvi falar desse tipo de ‘join’, mas suspeito que possa estar relacionado com uma ‘inner join’”. Ser aberto sobre o que não sabe é melhor do que estar confiante sobre algo incorreto; os entrevistadores desconfiam de pessoas que não sabem o que não sabem.

**DICA** Ao responder às perguntas de entrevista, seu instinto pode ser o de responder o mais rapidamente possível. Tente combater esse instinto; é muito melhor esperar para dar uma boa resposta do que responder mais ou menos. Com o estresse de uma entrevista, é difícil diminuir a velocidade com que fala, por isso pratique responder às perguntas antecipadamente para se sentir mais confortável.

Veja a seguir os tipos gerais de perguntas da entrevista técnica. Outra vez, verifique o Apêndice para ver exemplos de perguntas e respostas de entrevista.

- *Matemática e estatística* – as perguntas testam o quanto você entende de tópicos acadêmicos que são a base para o trabalho com ciência de dados. Elas incluem:
  - *Machine learning* – este tópico inclui o conhecimento de diferentes algoritmos de machine learning (médias  $k$  [ $k$ -means], regressão linear, floresta aleatória, análise de componentes principais, máquinas de vetores de suporte), diferentes métodos de utilização de algoritmos de machine learning (validação cruzada, boosting) e conhecimento geral em usá-los na prática (como quando certos algoritmos tendem a falhar).
  - *Estatística* – podem fazer perguntas puramente estatísticas, sobretudo

se seu trabalho seria em uma área que as responde, como em experimentos. Essas perguntas podem incluir testes estatísticos (como testes  $t$ ), definições de termos (como ANOVA e valor  $p$ ) e perguntas sobre distribuições de probabilidade (como encontrar o valor esperado de uma variável aleatória exponencial).

- *Combinatória* – esta área da matemática abrange tudo relacionado à contagem. Problemas lógicos nessa área incluem perguntas do tipo: “Se uma bolsa tem seis bolas de gude de cores diferentes nela, quantas combinações existem se tirar duas sem substituí-las?”. Essas perguntas têm pouco a ver com o trabalho de um cientista de dados, mas os entrevistadores às vezes acreditam que as respostas dão uma ideia da competência de resolução de problemas.
- *Bases de dados e programação* – essas perguntas testam sua eficácia nas partes baseadas em computador de um trabalho de ciência de dados. Elas incluem:
  - *SQL* – em quase todas as entrevistas de ciência de dados, haverá perguntas sobre como consultar bancos de dados com SQL. Esse conhecimento é necessário para a maioria dos trabalhos, e estar familiarizado com SQL indica que deve ser capaz de desempenhar sua nova função. Espere ser questionado sobre como escrever consultas SQL para dados de amostra. Você pode receber uma tabela de notas de alunos em várias disciplinas, por exemplo, e ser solicitado a encontrar os nomes dos alunos com a melhor pontuação em cada disciplina.
  - *R/Python* – dependendo da empresa, você pode ter que responder a perguntas gerais de programação escrevendo um pseudocódigo ou ter que resolver questões específicas usando R ou Python (qualquer linguagem que a empresa use). Não se preocupe se conhece R e a empresa usa Python (ou o inverso); normalmente, as empresas estão dispostas a contratar e treinar na nova linguagem no trabalho, uma vez que muito conhecimento é transferível. Espere perguntas que envolvam escrever código (por exemplo: “Como você filtraria uma tabela em R/Python para incluir apenas linhas acima do percentil 75 de uma coluna de pontuação?”).



- *Conhecimento de áreas corporativas* – essas perguntas dependem muito da empresa à qual está se candidatando. Elas são usadas para ver o quanto está familiarizado com o tipo de trabalho que a empresa faz. Embora pudesse aprender esse conhecimento no trabalho, a empresa preferiria se já o tivesse. Seguem algumas perguntas de exemplo utilizadas em diferentes setores:
  - *Empresa de comércio eletrônico* – qual é a taxa de cliques de um email? Como se compara com a taxa de abertura e como devem ser pensados de formas diferentes?
  - *Logística* – como otimizar filas de produção? O que se deve considerar no funcionamento de uma fábrica?
  - *Organizações sem fins lucrativos* – como uma organização sem fins lucrativos deve tentar medir o crescimento dos doadores? Como saber se muitos doadores não estão renovando?
- *Problemas lógicos complicados* – além de problemas relacionados à ciência de dados, podem fazer perguntas que envolvem a montagem de um quebra-cabeça na entrevista. Essas perguntas têm a finalidade de tentar testar sua inteligência e sua capacidade de pensar por conta própria. Na prática, as perguntas não fazem nada do tipo. A Google fez um estudo em massa (<http://mng.bz/G4PR>) e descobriu que essas perguntas não conseguiram prever a capacidade de um candidato no trabalho; serviram apenas para fazer o entrevistador sentir-se inteligente. Essas perguntas tendem a ser como “quantas garrafas de xampu existem em todos os hotéis dos Estados Unidos?”. Você pode fazer uma pesquisa no Google para ver se grandes empresas usam esses tipos de perguntas (e quais).

É difícil dizer exatamente quais dessas perguntas serão feitas e quanto tempo poderá passar nelas; esses fatores dependem muito da empresa e do seu entrevistador. Tente ao máximo manter-se calmo e confiante ao responder às perguntas, mesmo que não consiga responder a todas. Se o entrevistador estiver falando com você enquanto dá uma resposta parcial, ele pode pensar que está indo bem e estar disposto a apontá-lo na direção certa. Muitas vezes, as perguntas são pensadas para abranger tantos tópicos

que nenhum cientista de dados conseguiria responder a todas, então, logicamente, há algumas que não conseguirá responder.

## **Como entrevistar o entrevistador**

Toda vez que falar com uma nova pessoa durante a entrevista na empresa, ela terminará com: “Você tem alguma pergunta para mim?”. Essa pergunta é uma de suas únicas oportunidades de obter informações espontâneas sobre o trabalho, então use o tempo com sabedoria. Dá para descobrir mais sobre a tecnologia utilizada e o trabalho, bem como sobre a equipe. Suas perguntas mostram interesse sincero na empresa, então, vale a pena pensar nas perguntas antecipadamente. A seguir, alguns exemplos:

- “*Quais tecnologias você usa e como é o treinamento delas para pessoas novas?*”. Esta pergunta é excelente se a entrevista não tiver entrado em detalhes sobre os recursos de tecnologia. A resposta dará uma ideia se a empresa tem treinamentos formais ou espera que os funcionários aprendam por conta própria.
- “*Quem é o stakeholder da equipe e como é o relacionamento entre eles?*”. Esta pergunta é para saber quem toma as decisões do trabalho. Se o relacionamento for ruim, pode acabar virando um escravo das demandas desse stakeholder, mesmo que vá contra o que julgar certo.
- “*Como fazem o controle de qualidade no trabalho de ciência de dados?*”. Como a equipe não quer trabalhar com erros, deve haver algumas verificações no processo. Na prática, muitas equipes de ciência de dados não fazem verificações e culpam o criador do trabalho quando as coisas dão erradas. Esse tipo de local de trabalho é tóxico e deve ser evitado!

### **7.3.2 A entrevista comportamental**

A entrevista comportamental foi pensada para testar suas competências interpessoais e dar aos cientistas de dados da equipe uma ideia melhor de quem você é e qual é sua história. Embora as perguntas técnicas sejam colocadas em um ou dois blocos de tempo, as perguntas comportamentais podem aparecer durante toda a entrevista na empresa: em uma sessão de uma hora com alguém dos recursos humanos, em dez minutos nas perguntas de encerramento de um entrevistador técnico ou mesmo em uma conversa mais informal com um funcionário enquanto espera que outro chegue. Por isso, esteja preparado para responder a perguntas comportamentais a qualquer momento.

A seguir, exemplos de perguntas de entrevista a serem esperadas. Você

pode encontrar mais perguntas e respostas comportamentais na Seção A.4 do Apêndice:

- “*Fale um pouco sobre você*”. Esta “pergunta” é feita na triagem por telefone e pode ser refeita sempre que falar com uma pessoa nova. Mais uma vez, tente dar um resumo de um a dois minutos, mas, desta vez, adapte-o à pessoa com quem está falando.
- “*Qual seria um projeto em que trabalhou e o que aprendeu com ele?*” Esta pergunta pretende mostrar se você pode lembrar de um projeto da sua história e pensar bem sobre ele. Processou o que correu bem e o que não deu certo?
- “*Qual é sua maior fraqueza?*”. Esta pergunta dá nos nervos pois parece que, sob uma perspectiva da teoria dos jogos, é preciso dar uma resposta que mostra o mínimo de fraqueza possível. Na prática, o que a entrevista está testando é verificar se você compreende suas próprias limitações e se existem áreas que está tentando melhorar ativamente.

Observe que todas essas perguntas são muito abertas e não têm respostas certas ou erradas. Mas há maneiras de expressar-se que podem melhorar significativamente a maneira como suas respostas são recebidas.

Para a maioria dessas perguntas, especialmente aquelas sobre suas experiências, as respostas devem seguir uma estrutura geral:

1. Explique a pergunta com suas próprias palavras para mostrar que a entendeu.
2. Explique uma experiência na qual essa situação ocorreu, concentrando-se no motivo da existência do problema.
3. Descreva a atitude tomada para resolver o problema, bem como o resultado.
4. Faça um resumo do que aprendeu.

Considere esta solicitação: “Fale sobre uma situação em que tenha apresentado algo a um stakeholder e cujo retorno foi negativo”. Uma resposta poderia ser algo como: “Então, a pergunta é sobre uma situação em que desapontei alguém com meu trabalho [*1. Explique a pergunta de volta*]? Ocorreu algo no meu último emprego, onde tive de fazer um relatório sobre o crescimento do cliente. Nossa equipe estava lidando com

vários pedidos diferentes, então não tive muito tempo para me concentrar na solicitação de uma diretora. Quando entreguei um relatório em que passei apenas um dia, a diretora ficou muito desapontada [2. *Descrivendo o problema*]. Primeiro, pedi desculpas por não atender às expectativas dela; depois, trabalhei com ela para ver como poderíamos reduzir o escopo do pedido e ainda atender às suas necessidades [3. *Oferecendo uma solução*]. Com essa experiência, aprendi que é melhor avisar alguém logo se não conseguirá atender à sua solicitação para que ambos encontrem uma solução [4. *O que aprendeu*]”.

Uma coisa boa sobre perguntas da entrevista comportamental é que são todas similares, então dá para planejar as respostas! Se tiver três ou quatro histórias sobre trabalhar em uma situação desafiadora, lidar com colegas difíceis na equipe e falhas de gestão, você pode contar essas histórias na maioria das perguntas da entrevista. Essa abordagem é muito menos estressante do que tentar inventar uma história e improvisar na hora. Se tiver tempo, pode praticar contar essas histórias de carreira em voz alta a um amigo para saber como estruturá-las melhor.

Em quase todas as entrevistas, pedirão que descreva um projeto anterior, então vale a pena estar bem preparado. O projeto ideal para uma história abrange os pontos listados nesta seção: foi uma situação desafiadora em que superou a adversidade, especialmente colegas de equipe difíceis e, por fim, encontrou uma solução. Idealmente, a história também se encaixará no estilo de resposta de quatro passos.

Dito isso, muitas vezes é difícil encontrar uma história sobre um projeto que tenha sofrido uma reviravolta interessante, especialmente no caso de um aspirante a cientista de dados. Se não encontrar uma resposta, tente contar uma história simples sobre um projeto em seu portfólio. Mesmo uma resposta como: “Achei que seria interessante analisar um conjunto de dados incomum, então coletei, limpei e encontrei um resultado interessante que escrevi no meu blog” mostra ao entrevistador que você pode fazer uma análise.

A melhor técnica para responder a perguntas comportamentais é algo que muitas pessoas têm explorado. Você pode pensar sobre métodos até chegar ao nível do texto exato a ser usado e os segundos para gastar em cada

resposta. Dito isso, a maioria das técnicas não é específica à ciência de dados, então você pode ir mais longe lendo livros e artigos gerais sobre entrevista. Oferecemos recomendações na seção de recursos que consta no Capítulo 8.

## 7.4 Etapa 3: o estudo de caso

Se tiver se saído bem na entrevista na empresa, pedirão que complete um estudo de caso: um pequeno projeto que mostra à empresa sua ciência de dados na prática. Alguém na equipe de ciência de dados fornece um conjunto de dados, um problema vago para resolver com ele e um tempo definido para resolvê-lo. Podem pedir para resolver o problema durante a entrevista na empresa, com uma hora ou duas para trabalhar ou podem dar mais tempo, como um fim de semana, para realizar a tarefa em casa. Em geral, é possível usar as linguagens ou ferramentas de programação com as quais está mais familiarizado, embora seja possível que se limite às ferramentas da empresa. Quando o tempo acabar, você compartilhará seus resultados em uma breve apresentação ou discussão com um grupo de pessoas da equipe de ciência de dados. A seguir, alguns exemplos de estudos de caso:

- Com dados sobre emails promocionais enviados por uma empresa e dados sobre pedidos feitos, determinar qual das campanhas de email se saiu melhor e como a empresa deve comercializar de forma diferente no futuro.
- Com o texto de 20 mil tweets em que a empresa foi mencionada, agrupe os tweets em tópicos que acha que seriam úteis para a equipe de marketing.
- Um teste A/B caro foi executado no site da empresa, mas no meio do caminho, os dados deixaram de ser coletados de forma consistente. Com os dados do experimento, veja se algo pode ser concluído a partir deles.

Observe que, em cada exemplo de estudo de caso, o objetivo não é uma pergunta direta em ciência de dados. Perguntas como “qual campanha se saiu melhor?” fazem sentido em um contexto empresarial, mas não há um algoritmo “qual campanha se saiu melhor?” que possa ser aplicado. Estudos

de caso são úteis como ferramentas de entrevista porque requerem que você percorra desde o início de um problema até chegar a uma solução.

Dito isso, o que exatamente a empresa está buscando em um bom estudo de caso? Eles querem saber o seguinte:

- *Consegue pegar um problema vago e aberto e pensar em alguns métodos para solucioná-lo?* É totalmente possível não resolver o problema, mas desde que haja uma tentativa em uma direção razoável, assim você mostra que tem as competências técnicas e consegue fazer as coisas.
- *Consegue trabalhar com dados complicados do mundo real?* Os dados que receberá provavelmente precisarão de filtragem, criação de joins, engenharia de recursos e manipulação de elementos ausentes. Ao analisar um conjunto de dados complexo, a empresa está mostrando o tipo de trabalho que você faria.
- *Consegue estruturar uma análise?* A empresa quer saber se você analisa os dados de uma forma metódica, bem pensada ou se investiga coisas que não se relacionam com a tarefa em questão.
- *Consegue fazer um relatório útil?* Você terá de criar uma apresentação sobre seu trabalho e possivelmente alguns documentos, como no Jupyter Notebook ou relatórios em R no markdown. A empresa quer saber se você consegue fazer algo que seja útil para ela e estruturar uma narrativa útil.

A boa notícia é que as competências e técnicas necessárias para bons estudos de caso são exatamente as mesmas para projetos de um bom portfólio: analisar dados e uma pergunta vaga para produzir um resultado. É ainda melhor se fez um post de blog; isso simula a criação de uma apresentação para um estudo de caso de entrevista!

É bom levar em conta algumas pequenas diferenças entre um projeto de portfólio e um problema de estudo de caso:

- Em um estudo de caso, há um tempo limitado para fazer a análise. Esse tempo pode ser definido pelo calendário, como em uma semana a partir do dia em que receber os materiais ou definido pelas horas trabalhadas, como em não mais de 12 horas. Esse curto período de tempo significa

que é bom ser estratégico sobre como investir o tempo. Em geral, as etapas de limpeza e preparação de dados levam muito mais tempo do que os cientistas de dados esperam. Só reunir tabelas, filtrar caracteres malformados de strings e carregar os dados em um ambiente de desenvolvimento pode levar muito tempo e, normalmente, essa parte do trabalho não será impressionante para a empresa. Tente não se concentrar excessivamente na preparação dos dados da melhor forma possível, se significar ter muito pouco tempo para fazer uma análise.

- Outra diferença nos estudos de caso é que você é julgado com base na apresentação dos resultados: então, é bom ter uma apresentação bem feita com resultados realmente interessantes. O ato de criar uma apresentação pode parecer pouco interessante e menos importante do que fazer a análise em si; no entanto, muitos cientistas de dados adiam a criação da apresentação até o limite. Adiar a apresentação até o momento final é ruim porque pode ficar sem tempo ou pode descobrir que a análise não é tão interessante quanto esperava que fosse quando não há mais tempo para fazer alterações. Na medida do possível, comece a trabalhar na apresentação mais cedo e vá montando à medida que avança na análise.
- Uma orientação final sobre um estudo de caso é que você tem um público extremamente específico: o pequeno número de pessoas para quem apresentar. Com um projeto de portfólio, não se sabe de fato quem irá vê-lo, mas, com um estudo de caso, você pode hiperdirecionar sua análise. Se possível, quando receber o estudo de caso pela primeira vez, pergunte quem será o público-alvo da apresentação dele. Se o público for formado por cientistas de dados apenas, dá para fazer uma apresentação mais técnica, incluindo detalhes dos métodos de machine learning usados e por que você os escolheu. Se o público for de stakeholders, pegue leve com os componentes técnicos e foque mais em como suas descobertas afetariam as decisões da empresa. Se o público é uma mistura de cientistas de dados e stakeholders, tente incluir o suficiente de cada tipo dos detalhes de modo que se alguém do grupo dominar a discussão, você terá o bastante para argumentar.

A apresentação em si geralmente leva de 20 a 30 minutos para demonstrar



os resultados, com 10 a 15 minutos de perguntas do público sobre sua abordagem e seus resultados. É uma boa ideia praticar sua apresentação e planejar o que pretende dizer em cada parte. Praticar também ajuda a manter o tempo alocado; é bom não falar durante apenas 5 minutos ou durante 50 minutos consecutivos. Durante a parte de perguntas e respostas, muitas perguntas serão feitas sobre todos os tipos de tópicos. Podem surgir perguntas sobre um parâmetro em seu modelo e também sobre o efeito comercial na empresa do que você descobriu. Uma boa prática é parar para pensar sobre a pergunta antes de responder para que você respire e pense na sua resposta. No caso de não ter uma resposta segura, em geral é melhor dar uma versão de “não tenho certeza, mas...” e, depois, apresentar algumas ideias sobre como encontraria a resposta. Se possível, adicione um contexto relevante que você conhece.

## **7.5 Etapa 4: a entrevista final**

Quando o estudo de caso terminar, provavelmente durante a visita à empresa haverá uma última entrevista, a qual será feita com alguém que toma a decisão final, como o gerente da equipe de ciência de dados ou o diretor de engenharia. Dependendo de como a empresa conduz o processo de seleção, a pessoa pode estar preparada com informações sobre como se saiu nas primeiras partes do processo ou pode ser que não saiba sobre isso. O objetivo dessa entrevista é para que a pessoa final dê luz verde à sua contratação.

É difícil saber quais perguntas serão feitas durante a entrevista final, porque elas dependem muito da pessoa que o entrevistará. Uma pessoa com embasamento técnico pode focar em sua experiência técnica e em quais capacidades você tem. Já uma pessoa mais administrativa pode perguntar sobre resoluções de problema. Independentemente do tipo de pessoa que fizer a entrevista, deve-se definitivamente esperar perguntas que sejam do mesmo estilo daquelas na entrevista comportamental, por exemplo: “Como você lida com situações difíceis e com problemas?”. Para essas perguntas, seja aberto, honesto e sincero sempre.

## **Acompanhamento**

Após cada parte do processo de seleção, é possível que queira entrar em contato com as pessoas da empresa de alguma forma. Fazer o acompanhamento pode demonstrar gratidão às pessoas que conheceu, além de ter mais informações sobre como o processo está se desenrolando. Se fizer de maneira ruim, entretanto, entrar em contato pode parecer rude ou desesperado, além de comprometer suas possibilidades de conseguir a vaga. É possível utilizar qualquer um dos três métodos de acompanhamento a seguir, dependendo do local onde se encontra no processo:

- *Antes que alguém da empresa tenha entrado em contato*, se tiver enviado seu currículo, mas não tiver recebido retorno, não faça acompanhamento. A falta de retorno é um sinal de que a empresa não está interessada.
- *Após o contato, mas antes de se encontrar pessoalmente com alguém* – após a entrevista por telefone, só faça acompanhamento se não tiver certeza de seu *status* no processo. Você pode fazer acompanhamento com um email apenas se tiver passado o tempo em que a empresa disse que a próxima etapa aconteceria. Nesse caso, basta pedir uma atualização do estado.
- *Após o contato presencial* – você pode (mas de forma alguma precisa ou deve necessariamente) enviar um email com um breve agradecimento às pessoas que o entrevistaram. Se não receber retorno delas quando disseram que iriam mandar, também é possível enviar um email ao recrutador, solicitando uma atualização.

## 7.6 A proposta

Se tudo correr bem, dentro de uma ou duas semanas da entrevista final, você receberá uma ligação da empresa informando que ela fará uma proposta. Parabéns!

No Capítulo 8, damos mais detalhes sobre o que é uma boa proposta, como comparar propostas para diferentes vagas de ciência de dados e como pedir a uma empresa para melhorar a proposta.

Infelizmente, também é possível ficar desapontado; você pode não receber uma proposta da empresa. Depois de um momento para superar a perda do possível trabalho, é possível ver essa situação como uma oportunidade de

saber em quais áreas pode melhorar para a sua próxima entrevista. Se tiver feito apenas a triagem inicial por telefone, provavelmente é um sinal de que suas qualificações básicas não eram adequadas para a função específica. Nesse caso, considere ajustar as vagas para as quais se candidata. Se fez a entrevista na empresa ou estudo de caso, mas não mais do que isso, provavelmente há uma razão específica para não ser um bom candidato à empresa ou à função; tente deduzir se há algo em que poderia se concentrar na próxima entrevista. Se chegou às entrevistas finais, mas não conseguiu o emprego, normalmente significa que era um bom candidato à vaga, mas alguém foi um pouco melhor. Nesse caso, não há muito a fazer, mas continue se candidatando a vagas similares. Não entre em contato com a empresa para perguntar por que razão não foi contratado; é improvável receber uma resposta honesta, e a pergunta ainda pode ser vista como não sendo profissional.

## **7.7 Entrevista com Ryan Williams, cientista de decisão sênior da Starbucks**

Ryan Williams recentemente fez a transição de gerente de ciência de dados em uma empresa de consultoria de vendas e marketing, onde executava o processo de entrevista de ciência de dados. Agora na Starbucks, ele ajuda na tomada de decisões para o programa Starbucks Rewards, fazendo parte da equipe de análise e insights da empresa. Ele é bacharel pela Universidade de Washington, onde estudou estatística e economia. Antes de ingressar na Starbucks, sua carreira era focada no setor de consultoria.

### **O que fazer para se sair bem numa entrevista?**

O que mais conta, em geral, é apenas uma questão de preparação. Há todo um conjunto de competências que entra na entrevista e que é muito específico. Muitas pessoas pensam que podem só participar de uma entrevista e sua experiência falará por conta própria. Já me senti assim e vi outras pessoas assim também. Mas com perguntas como: “Fale sobre uma situação em que tenha tido problemas para se comunicar”, a menos que esteja realmente preparado, talvez se enrole e fique dando voltas. Existe um

conjunto de competências que são necessárias em uma entrevista, e a maneira de se preparar é ler as perguntas típicas que podem ser feitas. Não importa qual empresa está fazendo a entrevista, serão apresentadas algumas perguntas comportamentais, perguntas técnicas e perguntas sobre casos de negócios.

Então, pesquise as perguntas comportamentais mais comuns. Pesquise os tipos de perguntas técnicas que podem ser feitas e compreenda as perguntas de casos de negócio. A menos que esteja preparado para demonstrar sua experiência, não necessariamente você terá essa oportunidade em uma entrevista.

### **Como lidar com os momentos em que não sabe a resposta?**

Um caso que me recordo é de uma entrevista em uma das maiores empresas de tecnologia, e eles fizeram muitas perguntas complicadas de estatística. Cheguei ao ponto em que me perguntaram algo que era superacadêmico. Acho que me deram uma função de distribuição de probabilidade e perguntaram como eu usaria essa função de geração de momentos para encontrar a curtose de distribuição. Eu pensei: bem... talvez conseguisse responder algo assim na faculdade, mas definitivamente não consigo agora.

Quando dei minha resposta, o entrevistador ficou claramente desapontado. Poderia reclamar bastante sobre a pergunta, mas não fiquei contente porque sinto que era o tipo de pergunta mais de conhecimentos gerais do que algo que demonstraria meu raciocínio. Ter um trabalho é uma questão de saber usar suas competências para resolver coisas que não sabe. Não se trata da sua capacidade de entrar em uma sala sabendo tudo.

### **O que fazer quando receber um retorno negativo para sua resposta?**

Há o componente emocional, pois ficará insatisfeito com o fato de não ser capaz de responder à pergunta, mas é bom não deixar que isso acabe com sua entrevista. É natural ficar pensando no que poderia ter feito para resolver o problema em vez do que poderia fazer para responder à próxima pergunta. É preciso pensar rapidamente e continuar afiado para as próximas perguntas.

Eu diria que caso você se depare com perguntas difíceis assim, também precisa entrevistar bastante a empresa. Eles estão fazendo muitas perguntas, mas use os tipos de perguntas que estão fazendo para inferir se esta é uma empresa para a qual quer mesmo trabalhar. Para mim, uma empresa que pensa que é extremamente importante que um cientista de dados consiga derivar uma função geradora de momentos três vezes para obter a curtose não é necessariamente o tipo de ambiente em que sinto que será o melhor para se trabalhar.

### **O que aprendeu sobre entrevistas sendo um entrevistador?**

Sou muito mais atento aos tipos de perguntas que me fazem quando estou sendo entrevistado e também aos tipos de perguntas que faço em uma entrevista. Antes de começar a entrevistar, eu levava as perguntas da entrevista ao pé da letra. Ser entrevistado era como fazer um teste porque quando o entrevistador perguntava, eu pensava que era só dar a resposta correta. Não estava avaliando os tipos de perguntas sendo feitas, o que agora tenho muito mais consciência. O entrevistador está me fazendo perguntas que são muito específicas e complexas? Querem que eu liste muitos problemas de programação? Eles se preocupam com as coisas pelas quais me importo na ciência de dados?

### **Resumo**

- O processo de entrevista é similar na maioria das empresas para a área de ciência de dados.
- Para entrevistas na empresa, espere perguntas técnicas e comportamentais.
- Esteja preparado para fazer um estudo de caso em ciência de dados.
- Prepare-se ensaiando respostas para perguntas comuns de entrevista.
- Saiba o que perguntar sobre a empresa e a vaga durante a entrevista.

## CAPÍTULO 8

# A proposta: saber o que aceitar

Este capítulo abrange:

- O que fazer com a proposta inicial
- Como negociar a proposta de forma eficaz
- Como escolher entre duas opções “boas”

Parabéns! Você recebeu uma proposta de emprego em ciência de dados. É uma grande conquista, e deve tirar um tempo para saboreá-la. Foi preciso muito trabalho nos últimos meses ou até mesmo anos para chegar a esse ponto.

Este capítulo irá ajudá-lo a responder e a decidir sobre as propostas que receber. Embora provavelmente esteja muito entusiasmado, *não* deve dizer imediatamente: “Sim! Quando posso começar?”, ao receber uma proposta. Todos os empregos em ciência de dados não são criados da mesma forma. Os lugares para onde se candidatou foram bem selecionados, mas preocupações podem ter surgido na entrevista. Ou receberá um benefício que é uma exigência importante para você, como um plano de saúde bom para sua família, sendo necessário refletir sobre os detalhes da proposta. Mesmo que tenha certeza de que pretende aceitar a proposta, ainda não se deve dizer sim imediatamente: é melhor negociar! Ao ter recebido uma proposta, mas sem tê-la aceitado ainda, este é o momento de mais poder em suas mãos. Agora que o empregador finalmente encontrou alguém por quem está animado (você!), a empresa quer contratá-lo. O recrutamento é muito caro; demora muito tempo para que os recursos humanos e a equipe de ciência de dados avaliem os candidatos e os entrevistem, e cada semana sem a nova contratação é uma semana em que a empresa não se beneficia

do trabalho dessa pessoa (hipotética). Aproveite a oportunidade para pedir o que é importante para si, seja um salário maior, trabalhar em casa uma vez por semana ou um orçamento maior para participar de conferências.

## 8.1 O processo

Em geral, o processo da proposta é mais ou menos assim:

1. *A empresa avisa que fará uma proposta.* Ela avisa da proposta logo, para que você não aceite propostas de outra empresa antes.
2. *A empresa faz a proposta.* Por email ou por telefone (seguido de um email), a empresa informa os detalhes sobre salário, data de início e outras questões necessárias para sua decisão. Geralmente, também sugerem uma data-limite para aceitar ou recusar a proposta.
3. *Você dá uma resposta inicial.* Conforme será discutido na Seção 8.2, a menos que tenha certeza absoluta de que não deseja aceitar a vaga, recomendamos que expresse seu entusiasmo e peça alguns dias para pensar sobre a proposta em vez de dizer sim de pronto. Quando tiver a próxima conversa com a empresa, o processo de negociação começará (Seção 8.3).
4. *Negocie a melhor proposta que conseguir.* É possível receber uma resposta imediatamente ao negociar, mas, com frequência, a empresa precisa de um tempo para dizer se consegue oferecer uma proposta melhor.
5. *Decida se a proposta é boa o bastante e informe sua decisão final.*

## 8.2 Como receber a proposta

A ligação ou o email da proposta em geral é feito pelo gerente da contratação, recrutador ou pessoa de recursos humanos com quem tem conversado. Entretanto, independentemente da pessoa que estiver tratando da negociação, sua resposta deve ser a mesma.

Comece dizendo o quão feliz e entusiasmo está com a proposta. Se soar pouco entusiasmado em trabalhar lá, a empresa ficará preocupada com o fato de que, mesmo que a proposta seja aceita, você não ficará lá muito

tempo e não contribuirá com seu melhor trabalho.

**QUANDO ESTIVER MUITO DESAPONTADO** Embora geralmente recomendemos que comece a negociar depois de ter todos os detalhes por escrito e tenha tido alguns dias para pensar, se estiver muito desapontado com a proposta é recomendável discutir esse fato. Suponha que o salário é 25% menor do que o esperado. É possível começar dizendo algo como: “Muito obrigado. Estou muito entusiasmado com esta oportunidade e com o trabalho que faria na empresa Z. Contudo, gostaria de ser sincero e dizer que o salário está abaixo do que esperava. Sei que, na cidade de Nova Iorque, o valor de mercado para alguém como eu com mestrado e cinco anos de experiência está na faixa de X a Y. O que podemos fazer para chegarmos a uma proposta mais alinhada com essa faixa?”. Também poderia usar outra proposta maior recebida ou um salário atual maior como motivo para pedir mais. Deixe essa tática para quando o salário o faria de imediato recusar a proposta, não para quando a aceitaria, mas gostaria de negociar um salário 5% maior. Conseguir 20% a mais pode ser impossível, não importa quão bom negociador ou candidato você seja, e é melhor descobrir isso o quanto antes.

A empresa deve informá-lo que enviará os detalhes por email; caso contrário, pergunte a eles. Esse pedido tem duas finalidades:

- Ter tempo de ler tudo com calma e considerar o pacote total, sem precisar fazer anotações freneticamente durante uma ligação e tentar decifrar sua letra depois.
- Nunca considere a proposta como oficial até que ela seja feita por escrito. Na maioria das vezes, não haverá problema, mas é melhor evitar um mal-entendido por telefone e pensar ter aceitado determinado pacote de salário e benefícios, para depois descobrir que não é o caso.

A proposta recebida precisa incluir o título do cargo, salário, todas as opções ou ações oferecidas, além do pacote de benefícios. Se não incluir os detalhes necessários, como a explicação completa dos benefícios de seguro de saúde, também é possível pedir isso.

Por fim, deve estar especificado até que data é possível aceitar a proposta. O limite deve ser de, pelo menos, uma semana; se for inferior, peça uma semana. O melhor a fazer é estar confiante e dizer: “Preciso de alguns dias para considerar a proposta”. Se não conseguir ser tão seguro em suas convicções, é uma ótima hora para discutir a situação com outra pessoa, como um cônjuge, familiar ou um peixinho dourado da sorte. Contar com um tomador de decisões externo e algumas restrições ajudam nesse momento. Assim, o recrutador ou o gerente sabe que não pode pressioná-lo.

Às vezes, as empresas oferecem uma proposta “explosiva”, o que significa ter de responder em menos de uma semana; caso contrário, a



proposta será cancelada. Pode ser que o prazo não seja para responder nessa mesma ligação inicial, mas pode ser em apenas 24 horas. Normalmente, é possível estender o prazo dizendo algo como: “Entendo que ambos queremos que seja uma boa colocação. Escolher minha próxima empresa é uma grande decisão e preciso de uma semana para considerar a proposta com calma”. No pior cenário, por algum motivo, a empresa se recusará a esperar que a proposta seja totalmente avaliada. Se estiver nessa situação, é um grande sinal de alerta. Uma empresa que não esteja disposta a respeitar suas necessidades é uma empresa que não respeitará suas necessidades em outros momentos. Oferecendo um prazo curto, a empresa sabe que o está pressionando a agir rapidamente, o que causa mais ansiedade e pode levá-lo a tomar uma má decisão. Embora muitas pessoas estejam se candidatando a vagas na ciência de dados, as empresas ainda estão tendo dificuldade de encontrar bons candidatos. Então, uma empresa que trata os candidatos dessa maneira mostra que tem algo de estranho acontecendo nela.

Se estiver fazendo entrevistas em outras empresas, informe-as de que você recebeu uma proposta e até quando precisa responder. Se estiver na última rodada de entrevistas, essas empresas podem ser capazes de acelerar o processo para que possam fazer uma proposta antes de que decida sobre a outra proposta recebida. É totalmente normal dar atualizações a outras empresas ao receber outra proposta; elas gostarão disso. Quando entrar em contato com elas, reitere como está empolgado com eles e com o trabalho que faria lá. Algumas empresas podem não conseguir fazer alguma coisa, mas pelo menos você terá dado uma chance. Mais uma vez, como as empresas muitas vezes têm dificuldade para encontrar cientistas de dados para contratar, se encontrarem alguém de que gostaram, geralmente conseguem acelerar o processo.

## **8.3 A negociação**

Muitas pessoas odeiam negociar uma proposta de emprego. Uma das razões é a percepção de ser estritamente um jogo de soma zero: se ganhar, eles perdem. Outra razão é se sentir egoísta e ambicioso demais, sobretudo se a proposta for melhor do que seu cargo do momento. É fácil olhar para uma

situação e não sentir que merece mais dinheiro do que a empresa propõe inicialmente e que tem sorte de receber uma proposta. Ao estar nessa posição, você deve a si mesmo fazer o melhor para maximizar a proposta.

Você – sim, você – é o melhor candidato para a vaga. Todos enfrentamos a síndrome do impostor, mas a empresa vê em você o que ela quer para essa vaga. Como mencionamos no início do capítulo, é o melhor momento para negociar. E as empresas esperam que você negocie! É bom estar preparado. Um salário justo não está relacionado ao que tenha feito antes ou ao que a empresa pode propor inicialmente; em vez disso, depende do que as empresas estão propondo para as pessoas com seu conjunto de competências. Certifique-se de que seja pago tanto quanto seus colegas, que a compensação total combine com suas expectativas e que receba os benefícios que são mais importantes para você. Há a chance de aumentar 5% do seu salário em uma ligação de cinco minutos; dá para lidar com o desconforto por esse período.

A necessidade de negociar é especialmente forte na ciência de dados, que tem uma enorme disparidade salarial. Como a área é tão nova e como muitos tipos de funções acabam caindo na ciência de dados, não há normas claras para o que as pessoas fazem. Duas pessoas que têm o mesmo conjunto de competências podem ter salários totalmente diferentes quando alguém se chama de cientista de dados e a outra, de engenheiro de machine learning. Essas diferenças no salário aumentam à medida que as pessoas seguem para empregos novos e com salários maiores. A essa altura, há pouca correlação entre o montante que uma pessoa está ganhando e suas qualificações e competências.

### **8.3.1 O que é negociável?**

A primeira coisa que muitas pessoas pensam ao negociar uma proposta é o salário. Antes de chegar às rodadas finais de entrevistas, pesquise os salários não apenas para cientistas de dados em geral, mas também para o setor, a cidade em que a empresa está localizada e a empresa em particular, se ela já tem cientistas de dados. Embora a proposta possa ser um salto grande, lembre-se de que seu salário deve ser similar aos de seus colegas na empresa nova. Durante o processo de entrevista não é obrigatório informar

à empresa seu salário atual; inclusive em algumas cidades e estados dos EUA é ilegal que as empresas perguntem. Se pesquisar no Glassdoor e encontrar que o cientista de dados médio ganha o dobro que você, é ótimo.

Durante o processo de entrevista, se perguntarem seu salário atual ou quais são suas expectativas salariais, tente evitar responder, pois pode correr o risco de receber uma proposta mais baixa, pois a empresa sabe que pode atender às suas expectativas. Você pode tentar responder: “Neste momento, estou mais focado em encontrar uma vaga que seja adequada às minhas competências e experiência. Tenho certeza de que se for uma boa opção para ambos, poderemos chegar a um acordo sobre a compensação”. Se continuarem o pressionando, dizendo, por exemplo, que precisam saber para se certificarem de que não esperarão algo fora da faixa proposta, é possível responder: “Compreendo que querem ter certeza de que minhas expectativas correspondam à sua faixa de compensação. Qual é a faixa para esse cargo?”. Também pode dizer: “Preciso fazer mais algumas perguntas [ao gerente de contratação/cientista de dados sênior/qualquer pessoa que apareça no fluxo de entrevista com quem não estiver falando agora] para ter uma ideia melhor do que a vaga implica antes de dar uma expectativa realista”. Se a pessoa se recusar a falar da faixa e não seguir adiante sem um número, é um mau sinal. Se tiver absolutamente necessidade de dar uma resposta porque ama a empresa e tentou desviar várias vezes do assunto, diga algo como: “Embora seja bastante flexível, dependendo do pacote geral, sei pela minha pesquisa que o valor de X a Y é o padrão para alguém com minha experiência e formação”.

**Caitlin Hudon: sobre a síndrome do impostor**

Fazer grandes mudanças de carreira é um momento em que muitas pessoas – até mesmo cientistas de dados experientes! – sentem as dores da síndrome do impostor. A *síndrome do impostor* é o sentimento de dúvida sobre suas conquistas e a preocupação de parecer uma fraude. A negociação salarial é, em particular, complicada: é uma situação de risco elevado e um exercício literal de determinação do seu valor. Como você se valoriza certamente influencia sua tomada de decisão. Então, é importante combater os sentimentos da síndrome do impostor para que possa ser objetivo e assertivo sobre tudo o que está trazendo para a mesa.

A síndrome do impostor é especialmente comum na ciência de dados. Dependendo a quem perguntar, um cientista de dados é uma combinação de analista/estatístico/engenheiro/especialista em machine learning/visualizador/especialista em banco de dados/especialista de negócios, cada um com suas particularidades. Também é uma área em constante expansão, na qual pode haver muita pressão para se manter a par de novas tecnologias. Por isso, há pessoas com formações diferentes chegando a uma área nova com muitas aplicações, cujos limites não estão claramente definidos (causando, assim, lacunas inevitáveis no seu conhecimento sobre essa área como um todo) e em que a tecnologia está mudando mais rapidamente do que é possível acompanhar. Se parecer muito para uma só pessoa, é porque justamente é!

Isso nos leva à minha maneira de combater a síndrome do impostor, que consiste em focar na minha experiência única e na forma como me torna única, em vez de me comparar com um cientista de dados ideal e impossível de se alcançar. Para chegar lá, aceitei que nunca poderei aprender tudo o que há para saber na ciência de dados – nunca vou conhecer todos os algoritmos, todas as tecnologias, todos os pacotes interessantes ou até mesmo todas as linguagens – e tudo bem. (O bom de estar em uma área tão diversa e em crescimento é que ninguém saberá todas as coisas – e tudo bem também!)

Além disso, você e eu sabemos coisas e temos experiência que outros não têm. O conhecimento e a experiência que você tem se sobrepõem a de outros e também o distingue de outros, mas não é um subconjunto de outros. Foque na sua própria experiência única. Afinal de contas, você

fez para merecer, e é importante lembrar que isso o distingue de outros cientistas de dados.

Quando estiver no meio de negociações salariais, se se deparar com a síndrome do impostor, pare e pense em todas as competências que aprendeu, nos problemas que resolveu e em todos os grandes pontos positivos que traria para a futura equipe. Sua experiência é valiosa, e você deve ser bem pago por ela. Não tenha medo de pedir o que merece.

Não se sinta mal por pedir para receber o que considera razoável. O que é considerado razoável não deve ser definido pelo que fez antes, e sim pelo que você vai fazer no novo cargo. As empresas muitas vezes oferecem um valor mais baixo, antecipando que haverá negociação até o valor que esperam pagar pela nova contratação, portanto, é muito comum conseguir pelo menos um aumento de 5%.

Mas é possível negociar muito mais do que o salário. Primeiro, há outros benefícios monetários diretos, como bônus de assinatura de contrato, subsídio para mudança e ações. Um bônus de assinatura de contrato é mais fácil de a empresa pagar do que o mesmo valor de aumento de salário, já que se trata de um pagamento único. O mesmo se aplica a um subsídio de mudança; se precisar se mudar, pergunte sobre a política de mudança da empresa. Se a empresa for grande, é provável que exista um pacote de mudança-padrão, e pode até mesmo trabalhar com uma empresa de mudança em particular para ajudá-lo. Até mesmo as empresas pequenas podem ser capazes de oferecer algo; você não saberá se não perguntar.

Dá para ir além disso. Volte ao que pensou no processo da busca de emprego. O que é importante para você? Algumas coisas são difíceis de negociar; as opções de plano de saúde e previdência privada, por exemplo, são muitas vezes definidas pelos RH e são padrão para todos os funcionários. É melhor considerar esses fatores mais cedo no processo ao escolher para onde se candidatar. Para empresas, com exceção das startups, normalmente é possível encontrar comentários de funcionários no Glassdoor, que incluem informações sobre o pacote de benefícios. Mas há muitas coisas que você pode pedir agora, como:

- Um horário de trabalho flexível ou remoto.
- Uma revisão antecipada (seis meses x um ano), que poderia oferecer um aumento mais rápido.

- Benefícios educacionais.
- Um orçamento para ir a conferências.

Quando fizer isso, lembre-se das restrições da empresa. As organizações sem fins lucrativos, por exemplo, não têm muito espaço para negociar salários, mas podem ser flexíveis com horários ou férias. Uma empresa com uma equipe de dados já distribuída tem maior probabilidade de permitir trabalho em casa alguns dias por mês do que uma outra onde todos estejam no mesmo escritório. Certifique-se de que o que for pedido seja por escrito e cumprido. Negociar uma revisão antecipada e nunca a obter não é legal. Também tenha em mente que esses tipos de negociações não salariais são suscetíveis a ajustes apenas nas margens. Se o salário estiver muito abaixo do que quer, não há como essas mudanças tornarem a proposta aceitável.

### **8.3.2 Quanto negociar**

A melhor negociação é com uma proposta concorrente, a qual permite que a empresa saiba que existe um mercado para seus serviços a um preço mais elevado e que há outro lugar para ir. Se se encontrar com duas propostas, é melhor dizer algo como: “Prefiro trabalhar com vocês, mas a empresa ABC está me oferecendo muito mais. Vocês conseguem cobrir a proposta?”. Não minta; se tentar falar isso a ambas as empresas, facilmente o fato voltará para assombrá-lo.

Outro fator é seu trabalho atual. A maioria das empresas sabe que você não quer uma redução salarial. Se estiver relativamente feliz ou tiver melhores benefícios em sua empresa, use esse fato. Mesmo se a proposta tiver um salário maior, lembre-se também de calcular o valor recebido pelos seus benefícios. Suponha que seu empregador atual ofereça 3% de previdência privada, e a empresa da proposta não ofereça nada. Sua empresa atual aumenta seu salário em 3% em comparação à nova empresa com a proposta. Ou talvez sua empresa pague plano de saúde completo, e, na da proposta, terá de pagar 200 dólares por mês pelo plano de saúde para sua família. Esse fato pode ser especialmente útil se estiver fazendo a transição para a ciência de dados vindo de uma área com salários mais baixos, de modo que seu salário aumentará muito, mas estará perdendo um benefício que recebe no seu trabalho atual. É possível trazer esses pontos

mencionados como as razões pelas quais está pedindo um salário maior.

Ao atribuir um valor às suas competências, considere como sua formação única se adapta ao seu novo trabalho. Não precisa necessariamente ser um cientista de dados experiente. Seria ótimo ser uma das três pessoas a terminar o doutorado em inteligência artificial em Stanford nesse ano. Mas suponha que a proposta é ser cientista de dados que ajude a área de vendas e que você trabalhou nessa área. Ter conhecimento da área é uma enorme vantagem que os cientistas de dados, mesmo mais experientes, podem não ser capazes de oferecer ao cargo. Quanto mais parecer que o cargo foi feito para você ao ver a publicação da vaga, mais vantagem terá.

Às vezes, pode ser difícil de enxergar todos os aspectos de uma proposta de trabalho. Se precisar se mudar, por exemplo, considere os custos de mudança e o custo de vida. Um salário de 90 mil dólares em Houston rende muito mais do que um salário de 95 mil dólares em Nova Iorque. Coloque na conta os benefícios adicionais. Coisas como um bom plano de saúde ou de previdência privada podem valer muito a pena. Em geral, não perca de vista todos os componentes do salário.

Se tiver negociado e receber tudo o que foi solicitado, a empresa espera que você aceite a proposta! Inicie o processo de negociação somente se for aceitar a proposta caso as suas condições sejam atendidas. Caso contrário, por que ter todo o trabalho? Uma razão poderia ser ter outra proposta que gostaria de aceitar, e quanto melhor uma proposta for, mais é possível negociar a que você quer. Essa tática é extremamente arriscada, entretanto, é bom tomar cuidado para não brincar com fogo. Mesmo que nunca trabalhe na empresa que está sendo usada para aumentar a outra proposta, os funcionários que o entrevistaram ou fizeram a proposta podem se lembrar.

É muito raro que uma empresa cancele uma proposta que foi negociada. Se acontecer, é um sinal de que é definitivamente melhor não trabalhar lá! É totalmente normal negociar com respeito; rejeitar um candidato por essa razão é um enorme sinal de alerta. Leve isso como uma situação ruim que foi evitada.

## **Um breve manual sobre RSUs, opções e planos de compra de**



**ações de funcionários**

Esta seção é intencionalmente designada de “um breve manual”, e não como “visão geral exaustiva”. Recomendamos fortemente que pesquise mais sobre o que está contido na sua proposta.

Em geral, todos os itens a seguir valem para um período de quatro anos com uma carência de um ano. *Carência de um ano* significa que se sair da empresa antes de um ano, não ganhará nada. Em vez disso, receberá o valor de um ano todo de uma vez no seu aniversário de um ano na empresa. Depois disso, geralmente recebe-se a cada trimestre ou mês nos três anos seguintes:

- *Unidades restritas de ações (RSUs – Restricted Stock Units)* – você verá uma quantidade de RSUs dada a você em uma proposta. Se tomar o valor da ação no momento da proposta e dividir o montante monetário por ele, obterá o número de ações ao longo do tempo à medida que foram investidas. Se as ações subirem em valor, sua compensação também sobe. Quando as ações são investidas, você as receberá como títulos naquele momento; sua empresa geralmente reserva algum valor para impostos na forma de ações.

Suponha receber uma proposta de 40 mil dólares, investindo em mais de quatro anos com uma carência de um ano. Nesse momento, as ações estão sendo negociadas a 100 dólares por ação; então, receberá 100 ações após um ano e 25 ações por trimestre, depois. Depois de trabalhar lá por um ano, você receberá 65 ações (35 retidas para impostos). A essa altura, é possível fazer o que quiser com suas ações, desde que você cumpra com as regras da empresa. (Via de regra, você pode vender as ações somente durante determinados períodos, por exemplo.) Se estiver trabalhando para uma empresa privada, os investimentos funcionam da mesma maneira, mas em geral não poderá vender suas ações até que a empresa entre na bolsa ou seja adquirida.

- *Opções de ações* – opções dão a você a “opção” de comprar uma determinada quantidade de ações a um preço de exercício especificado, que é geralmente o justo valor de mercado das ações no momento em que a opção é concedida. Se a ação for negociada mais tarde com preço superior a esse, é uma excelente notícia; se puder comprar um título a 10 dólares por ação e estiver negociando a 30 dólares, ganhará 20 dólares só entre comprar e vender as ações! No entanto, se o título não for negociado ao preço do exercício, não vale a pena; não é recomendável exercer uma opção para comprar um título no valor de 10 dólares se puder comprá-lo no mercado por 5 dólares.

As opções na verdade são ótimas para os funcionários mais antigos em uma empresa que vai crescendo, mas também podem ser como bilhetes de loteria. Desde que sua empresa permaneça privada, as opções têm valor limitado, e, mesmo se a empresa entrar na bolsa, o título pode acabar sendo negociado abaixo do valor de opção. Ao

contrário das RSUs, que geralmente valerão alguma coisa, as opções podem nunca valer a pena o dinheiro.

- *Planos de compra de ações dos funcionários (ESOPs – Employee stock purchase plans)* – esses planos permitem comprar ações da empresa com desconto. Você contribui para o plano por meio da dedução na folha de pagamento e, em seguida, chega a uma data de compra na qual seu dinheiro é usado para comprar ações da empresa. Esses planos oferecem dois benefícios:
  - Desconto sobre o preço das ações, que pode ser de até 15%.
  - Provisão de direito de ressarcimento na venda de ações. Se o preço tiver subido entre o início do período da proposta e a data de compra, terá de pagar o preço mais baixo.

Suponhamos que, no início do período da proposta, as ações da empresa valessem 10 dólares. Você contribui com 9 mil dólares ao longo de um ano e, então, atinge a data de compra. Com desconto de 10%, você começa a comprar 1.000 ações (US\$ 9.000/US\$ 9). Se o preço das ações for agora de US\$ 20, essas 1.000 ações valem 20 mil dólares, o que significa que poderá vendê-las e obter 11 mil dólares de lucro!

## 8.4 Táticas de negociação

Agora que já passamos pelo panorama geral, vamos ver algumas dicas de negociação específicas:

- *Lembre-se de começar mostrando que está agradecido e empolgado.* Esperamos que ambos estejam! É bom fazer a empresa sentir que você está ao lado dela e trabalhando para chegar a uma solução. Não sentir que estão trabalhando em conjunto é um sinal de alerta para não trabalhar lá.
- *Esteja preparado.* Antes da ligação, prepare anotações sobre exatamente o que quer alterar na proposta e de que compensação total gostaria. No calor do momento, é extremamente fácil jogar com um número mais baixo do que queria para deixar a outra pessoa feliz. Ter anotações o ajudam a evitar essa situação.

- *Escute o que a outra pessoa está dizendo.* Se enfatizarem que não negociam salário, mas a compensação é importante para você, pergunte sobre mais opções ou um bônus de assinatura de contrato. Tente trabalhar com eles para encontrar uma solução em vez de ser inflexível. Embora as negociações possam parecer um jogo de soma zero, não precisa ser! Algo muito importante para você pode não ser tão importante para a empresa, por exemplo, então é fácil para a empresa oferecer. Lembre-se de que a empresa quer que você fique feliz para ter sucesso no seu cargo.
- *Não pareça que está focado somente no dinheiro* (mesmo que esteja). Essa atitude pode soar mal. É melhor mostrar que está motivado pelo trabalho que está fazendo, pela missão na empresa e seus novos colegas. Parecer que só está ali pelo dinheiro pode soar muito mal para o recrutador, e é mais provável que se preocupe que você trocará de empresa se surgir um salário mais alto.
- *Tente manter um foco comum.* Em vez de dizer “preciso de mais opções”, por exemplo, diga “estou muito entusiasmado com as perspectivas de crescimento a longo prazo da sua empresa e com a forma como posso ajudá-los a obter sucesso. É por isso que estou interessado em receber mais investimento com algumas opções adicionais”.
- *Agrupe seus desejos em vez de ir ponto a ponto.* Dessa forma, o recrutador tem a visão geral do que você quer, em vez de sentir que cada vez que um problema é resolvido, outro pode surgir. Dito isso, pense nos substitutos que está disposto a aceitar. Talvez peça um salário maior, mais opções e poder trabalhar em casa um dia por semana, mas se a empresa não puder cobrir o salário, dá para pedir um bônus de assinatura de contrato. Se possível, liste também a importância de cada item para você.
- *Evite autodepreciação.* Leu a parte anterior neste capítulo sobre quão bom você é? Leia-a novamente. Não se desvalorize dizendo algo como “sei que não trabalhei como cientista de dados antes” ou “sei que não tenho doutorado, mas...”. Você tem exatamente o que a empresa está

procurando, ou ela não teria feito uma proposta!

Você pode encontrar muito mais conselhos e pesquisas sobre como negociar efetivamente em vários artigos e livros, e compartilhamos alguns de nossos favoritos no Apêndice.

**COMO NEGOCIAR SENDO UMA MULHER** Por algum tempo, a teoria comum de pesquisa foi a de que “as mulheres não pedem”. Essa teoria foi uma das razões dadas pelas quais mulheres recebem menos do que os homens: afirma que as mulheres não negociam propostas de emprego nem pedem aumentos. Pesquisas recentes, no entanto, mostraram que, pelo menos em algumas áreas, as mulheres negociam tanto quanto os homens; elas só têm menos probabilidade de serem bem-sucedidas. Infelizmente, existe preconceito. Algumas das táticas que discutimos nesta seção, como ter um foco comum (usando *nós* em vez de *eu*, por exemplo), são especialmente benéficas para as mulheres.

## 8.5 Como escolher entre duas “boas” propostas de trabalho

Receber duas (ou mais!) boas propostas é um ótimo problema para se ter! Mas ainda é preciso fazer uma escolha. Naturalmente, várias propostas nem sempre acarretam esse problema. Às vezes, a preferência maior por um trabalho facilita a decisão. Mas o que fazer se não for esse o caso?

O primeiro passo é voltar à Seção 8.3. Negocie! Se gostar de uma empresa mais do que de outra, mas houver algum empecilho para você, veja se a empresa pode mudar essa situação. Seja honesto quanto ao que está buscando. Como discutido anteriormente, essa proposta concorrente é uma grande vantagem para se ter e facilita conseguir o que está buscando.

Em seguida, pode pedir para se reunir novamente com o gerente da contratação ou com futuros colegas de trabalho para obter mais informações. Talvez não tenha perguntado o número médio de reuniões semanais, com quantas equipes estaria trabalhando, como as solicitações são priorizadas ou qualquer outra coisa que seria útil para avaliar se essa empresa é um lugar onde gostaria de trabalhar.

Quando estiver considerando uma proposta, leve em conta o longo prazo. Pode haver uma preocupação a curto prazo: se tiver uma dívida elevada de estudante ou estiver começando uma família logo, pode ficar tentado a aceitar a proposta que pague melhor, para poder começar a pagar suas

dívidas ou a guardar dinheiro. Se tiver a sorte de poder pensar para além das preocupações financeiras imediatas, foque na maximização do potencial a longo prazo. Que tipo de trabalho estaria fazendo em cada empresa? Se um deles oferecer treinamento que pode proporcionar subir dois degraus para o cargo seguinte e o outro exige coisas que aprendeu em seu primeiro dia no bootcamp, não deve se atirar na última opção apenas porque a empresa tem um pacote de férias melhor. Lembre-se de que os salários na ciência de dados podem saltar de um cargo a outro em até 30 ou 40%. Se seu currículo está com algumas lacunas, e esse trabalho é uma boa chance de colocar um nome de prestígio nele, talvez tenha que aceitar um emprego com um salário um pouco menor por um ano ou dois.

Por fim, não tenha medo de deixar que fatores menores o influenciem. Um trabalho envolve um deslocamento mais curto ou um escritório mais espaçoso? Se as empresas atenderem aos seus critérios mínimos e forem similares quanto aos fatores de decisão mais importantes, você pode começar a considerar os fatores menores.

Em situações como essa, considerar diversas escolhas de vida pode ser bom. Melhor aceitar o trabalho na startup emocionante, mas arriscada, ou seguir no seu trabalho mediano em uma empresa governamental? Melhor aceitar a proposta em que você gerencia pessoas ou continua a trabalhar como independente?

Muitas vezes, não há uma maneira objetiva de decidir qual rota é a melhor. Não é possível saber se o novo emprego seria uma boa opção até aceitar de fato a proposta. Também não tem como saber se irá gostar de gerenciar até se tornar um gerente. É um fato estressante da vida.

Se estiver se debatendo com essas decisões, pode ajudá-lo pensar que só pode fazer o melhor que consegue com a informação que tem em mãos. Não se pode prever o futuro! Se tomar uma decisão e se sentir infeliz com o resultado, dá para seguir em frente. É possível sair de um emprego, voltar depois de mudar de cidade ou voltar a ser independente. A vida é complicada, mas é possível aprender muito até mesmo com caminhos que não levam aos resultados desejados.

Quando tomar uma decisão, recuse as demais propostas graciosamente. É bom ser educado, e o mundo da ciência de dados é pequeno. Não se rejeita

uma proposta dizendo: “Não posso acreditar que me ofereceram essa proposta tão baixa; que empresa terrível e completamente antiética!”, vindo depois a descobrir que a pessoa com quem gritou é o gerente de contratação do seu trabalho dos sonhos cinco anos depois.

## **8.6 Entrevista com Brooke Watson Madubuwu, cientista de dados sênior da ACLU**

Brooke Watson Madubuwu é cientista de dados sênior da American Civil Liberties Union (ACLU), onde presta apoio quantitativo a equipes de litígio e advocacia em questões relacionadas aos direitos civis. Ela tem mestrado em epidemiologia e trabalhou anteriormente como cientista de pesquisa.

### **O que se deve considerar além do salário ao avaliar uma proposta?**

Conhecer suas prioridades antes de entrar no processo pode ajudá-lo. É fácil comparar salários, mas também é bom comparar a experiência cotidiana ou mensal do trabalho em cada lugar. Tenho tendência a pensar nos aspectos do trabalho em três categorias: estilo de vida, aprendizagem e valores.

O estilo de vida é sobre como o dia a dia do trabalho interage com outras partes da vida. Onde irei morar? Posso trabalhar remotamente? Trabalharei à noite e aos fins de semana? Consigo viajar? Terei de viajar? Quais são as opções de plano de saúde, cuidados infantis e investimento de aposentadoria disponíveis para mim? Tenho a flexibilidade de que preciso para cuidar da minha família? Esses são aspectos referentes ao estilo de vida.

Há também aspectos referentes à aprendizagem. Crescerei nesse cargo? Quais sistemas existem para garantir que irei crescer? Estou entusiasmado em aprender com meu chefe e com a equipe?

Por último, penso nos valores e na missão da empresa ou organização e da equipe específica. A organização e a equipe trabalham em prol de algo que esteja alinhado com meus valores? A equipe valoriza a inclusão? É um produto ou uma equipe que quero continuar a formar? Cada uma dessas três caixinhas pode ter níveis diferentes de importância em momentos distintos.



Negocie uma mudança no título antes em um cargo anterior, e é possível negociar coisas como recursos para formações e conferências, dias de trabalho em casa, suporte a viagens ou ações. A ciência de dados é uma área vasta que está sempre avançando, por isso, penso que também é importante reservar algum tempo do seu dia para continuar a desenvolver suas competências.

### **Quais são algumas formas de se preparar para negociar?**

No passado, sentia-me muito desconfortável em falar sobre meu próprio valor e a defendê-lo. É algo que tenho tentado desaprender ao longo da minha carreira. Pode ser muito útil ter aliados por perto. Pratique a negociação com um amigo próximo e deixe que ele faça seu papel também. Muitas pessoas, eu mesma incluída, acham muito mais fácil defender nossos amigos do que nós mesmos. Ouvir a forma como alguém descreve seus pontos fortes e suas necessidades pode ser muito motivador e ajuda bastante a colocar-se nessa mentalidade. Pergunte-se: “Como falaria de alguém que amo, sabendo que essa pessoa é uma opção excelente para este cargo e que gostaria que fosse bem-sucedida?”. É assim que deve falar sobre si mesmo.

### **O que fazer se tiver uma proposta, mas ainda estiver à espera de outra?**

Se tiver passado por todo ou pela maior parte do processo de entrevista de uma empresa pela qual está esperando, não dói mencionar a eles que há outra proposta. Também é possível ganhar mais tempo com o empregador da proposta pedindo uma semana ou duas para rever os detalhes e discutir a proposta com sua família. Negociar mais, fazer perguntas detalhadas sobre os benefícios e pedir para falar com a equipe sobre o trabalho ou a cultura são todas ações que se pode tomar para servir à finalidade dupla de coletar mais informação sobre o cargo e ganhar mais tempo.

Se ainda estiver muito no início do processo de entrevista de outra empresa, talvez não consiga acelerá-lo, e terá de comparar a proposta com sua situação atual. Se tiver recém se graduado ou estiver sem trabalho, talvez não tenha o luxo financeiro da espera. Se estiver empregado

atualmente e ainda não estiver interessado pela proposta, pode valer a pena permanecer no cargo atual até encontrar uma opção melhor. A minha primeira proposta de ciência de dados não era uma opção excelente e, embora estivesse muito interessada em fazer a transição para a área, valeu a pena esperar por uma melhor.

## **Qual é seu último conselho para os cientistas de dados júnior e aspirantes?**

Gostaria de aconselhar os cientistas de dados iniciantes e aspirantes a manterem uma mente aberta sobre os títulos. Acho que o título de cientista de dados exerce esse fascínio nas pessoas, especialmente para aqueles vindos dos cursos de ciência de dados que não existiam quando eu fazia faculdade. Eles podem sentir que, se não conseguirem um emprego como cientista de dados logo após a graduação, então falharam. Porém, as funções do dia a dia de um cientista de dados podem fazer parte de um analista de dados, um pesquisador ou um de muitos outros tipos de cargos, que podem ser um terreno muito grande para aperfeiçoar suas competências. Desenvolvi minhas habilidades de programação trabalhando como assistente de pesquisa e, mais tarde, como cientista de pesquisa, durante anos antes de alguém me chamar de cientista de dados, e o trabalho era tão interessante quanto envolvente. Mesmo o trabalho de inserção de dados como graduanda moldou meu pensamento sobre a forma como as decisões de coleta de dados informam as possibilidades analíticas. Nenhum trabalho é muito pequeno se estiver disposto a aprender ao longo do caminho.

## **Resumo**

- Não aceite uma proposta imediatamente: receba os detalhes por escrito e peça um tempo para considerá-la.
- Negocie, negocie, negocie! Você pode pedir um salário mais alto ou outros benefícios monetários, mas não se esqueça de coisas como um horário de trabalho flexível e recursos para participar de conferências.
- Quando estiver pesando duas boas propostas, lembre-se de considerar o

potencial a longo prazo de cada uma delas, não apenas o salário inicial.

## **Recursos dos capítulos 5–8**

### **Livros**

*I Will Teach You to be Rich*, 2. Ed., por Ramit Sethi (Workman Publishing Company)

Esta franquia é um blog, um livro, vídeos e várias outras mídias, todas úteis para conseguir um emprego. Sethi tem recursos incríveis sobre como pensar uma entrevista e dar respostas que serão bem recebidas. Esses recursos também têm ferramentas para negociação salarial, pedir um aumento e outras conversas complicadas que podem acontecer ao defender seu valor.

*What Color Is Your Parachute? 2020: A Practical Manual for Job-Hunters and Career-Changers*, por Richard N. Bolles (Ten Speed Press)

É um livro para encontrar um novo emprego. Muito parecido com os capítulos 5 a 8 desta publicação, ele inclui entender o que se quer em um trabalho, fazer uma busca, montar o currículo, entrevistas e negociação. Oferece uma boa visão de como pensar amplamente sobre encontrar um emprego, embora nada nele seja específico a empregos técnicos (muito menos de ciência de dados).

*Cracking the Coding Interview*, por Gayle Laakmann McDowell  
(CareerCup)

*Cracking the Coding Interview* é um livro dedicado a ajudar os engenheiros de software a conseguir emprego, incluindo quase 200 perguntas e respostas de entrevistas. Embora muitas das questões não sejam relevantes para a ciência de dados, muitas delas são, especialmente se estiver procurando um trabalho como engenheiro de machine learning que tenha muita ênfase em programação.

## **Posts de blog e cursos**

*“Advice on applying to data science jobs,”* por Jason Goodman

<http://mng.bz/POlv>

Este é um excelente post de blog sobre as lições aprendidas ao se candidatar para empregos em ciência de dados e descobrir como são generalizados. O post abrange muitos dos conceitos que estão nos capítulos 5 a 8, mas com a perspectiva pessoal do autor sobre como se sentiu durante o processo.

*“How to write a cover letter: The all-time best tips,”* por Muse Editor

<http://mng.bz/Jzja>

Embora o Capítulo 6 forneça orientação sobre como escrever uma carta de apresentação, é bom ter várias perspectivas. Este post de blog trata de tópicos que não estão neste capítulo, por exemplo: como não se preocupar em se elogiar.

*“Up-level your résumé,”* por Kristen Kehrer



[https://datamovesme.com/course\\_description/up-level-your-resume](https://datamovesme.com/course_description/up-level-your-resume)

Nossa entrevistada do Capítulo 6 criou esse curso pago para ajudar os aspirantes a cientistas de dados a otimizarem seus currículos e cartas de apresentação para conseguir passar pelos sistemas de acompanhamento de candidatos (também conhecidos como robot screeners) e recrutadores. Se não estiver conseguindo muitas entrevistas de primeira rodada, este curso pode ser para você.

*“How to quantify your resume bullets (when you don’t work with numbers)”*, por Lily Zhang

<https://www.themuse.com/advice/how-to-quantify-your-resume-bullets-when-you-dont-work-with-numbers>

Se estiver se debatendo com nossos conselhos sobre como quantificar suas experiências no currículo, este post pode ajudar. Abrange três formas de quantificar sua experiência: intervalo, frequência e escala.

*“How women can get what they want in a negotiation,”* por Suzanne de Janasz e Beth Cabrera

<https://hbr.org/2018/08/how-women-can-get-what-they-want-in-a-negotiation>

Este post de blog da *Harvard Business Review* abrange como as mulheres podem superar os preconceitos inerentes contra elas ao negociar salários e propostas de emprego. É um tópico valioso para as mulheres pensarem, pois o valor de negociar de forma mais eficaz durante uma carreira cresce com o tempo.

*“Ten rules for negotiating a job offer,”* por Haseeb Qureshi

<https://haseebq.com/my-ten-rules-for-negotiating-a-job-offer>

Este post de blog trata com profundidade sobre como negociar com sucesso uma proposta de emprego. Quando se encontrar na fase de se candidatar a um emprego, ter a informação certa pode significar uma diferença de milhares de dólares; por isto, leia este post ao se aproximar dessa fase.

## PARTE III

# Adaptação na área de ciência de dados

Começar seu primeiro emprego de ciência de dados é um grande feito, mas é apenas o início da sua carreira na área. Trabalhar em uma empresa como cientista de dados é bastante diferente de fazer ciência de dados como um hobby ou como parte de um curso. Pode ser necessário aprender diversos conceitos, desde a rotina na empresa à maneira apropriada de colocar o código em produção. A enorme diferença entre a sua expectativa de como será o cargo e como é de fato o trabalho pode ser um choque! A Parte III do livro pretende prepará-lo para isso. Ao ler esta parte, saberá o que esperar de um trabalho em ciência de dados e estará mais bem preparado para se destacar.

O Capítulo 9 trata sobre os primeiros meses no trabalho, desde os primeiros dias, quando pode se sentir totalmente perdido, passando pelo período de adaptação à medida que aprende mais sobre o cargo, seus colegas de trabalho e os dados. O Capítulo 10 contém um guia para a criação de boas análises (uma boa parte da maioria das funções da ciência de dados), fazendo e executando um plano no início. O Capítulo 11 discute modelos de machine learning e a colocação dos mesmos em produção, introduzindo conceitos como testes de unidades (unit testing) que são essenciais para funções de ciência de dados mais baseadas em engenharia. O Capítulo 12 é um mergulho profundo na tarefa extremamente relevante de trabalhar com stakeholders, que muitas vezes faz parte do trabalho de ciência de dados e no qual muitos enfrentam dificuldades.

## CAPÍTULO 9

# Os primeiros meses de trabalho

Este capítulo abrange:

- O que esperar das suas primeiras semanas como cientista de dados
- Como se tornar produtivo construindo relacionamentos e fazendo perguntas
- O que fazer caso acabar se encontrando em um ambiente de trabalho ruim

Neste capítulo, vamos explicar o que esperar dos primeiros meses e como utilizá-los para se preparar para obter sucesso. Esse período terá um impacto muito grande sobre a forma como o trabalho será conduzido; essa é a sua oportunidade de criar um sistema e uma rede de apoio que lhe permitirá ser bem-sucedido. Embora cada trabalho de ciência de dados seja diferente, alguns padrões e princípios amplos aplicam-se a qualquer trabalho.

Quando começar a trabalhar, é instintivo querer fazer o máximo possível. Lute contra esse instinto. É preciso ter certeza de que não está apenas realizando tarefas, mas fazendo-as da maneira correta. Quando se está começando em um novo trabalho, é o momento mais fácil de fazer perguntas sobre como algo deve ser feito, pois não se espera que você saiba os processos da nova empresa. Ocasionalmente, os gerentes se esquecem de que você não tem o conhecimento institucional que seu antecessor pode ter tido, então, talvez fique responsável por algo que não faz sentido para você. É possível que consiga fingir nas primeiras tarefas, mas será muito melhor fazer perguntas logo e descobrir como abordar seu processo de trabalho.

### 9.1 Primeiro mês

Seu primeiro mês em uma empresa será diferente de seu primeiro mês em

outra empresa. As pequenas e grandes empresas o receberão de maneiras quase opostas. A Figura 9.1 compara o que se pode esperar em duas empresas: uma enorme, com muitos de cientistas de dados e outra com nenhuma ou quase nenhuma equipe de ciência de dados. (No Capítulo 2, esses exemplos seriam a MTC e a Seg-Metra, respectivamente.) Esses dois exemplos destacam os fins de um espectro, mas a empresa para a qual está trabalhando provavelmente se situará entre elas.



*Figura 9.1 – A integração em uma grande empresa é como passar por uma linha de fábrica, enquanto em uma pequena empresa é mais direto. (Emojis do Twitter do projeto Twemoji.)*

### **9.1.1 Integração em uma grande empresa: uma máquina bem lubrificada**

Você é uma das dezenas de pessoas que começam esta semana. Recebeu um email na semana anterior dizendo aonde ir, quando chegar e o que precisa levar. Agora, começa um processo formal de integração de vários dias com pessoas de diferentes departamentos. Você recebe seu computador e passa pelo processo de configuração. Escuta apresentações sobre a cultura da empresa, políticas de recursos humanos e como a empresa é organizada. Tudo funciona como um relógio; a empresa já integrou milhares de funcionários.

Quanto à ciência de dados, receberá ajuda para configurar seu ambiente de programação. É provável que haja uma checklist ou documentação extensa sobre tudo que é preciso fazer para acessar os dados. Há também um repositório central de relatórios antigos e documentação dos dados para ler e absorver. Ninguém espera que saiba fazer muito imediatamente; embora seus colegas estejam entusiasmados por tê-lo na equipe, eles sabem

que será necessário tempo para você se adaptar. Espera-se que leve algumas semanas para passar por todo o treinamento e obter acesso aprovado nos sistemas. Pode ser que se sinta frustrado por demorar tanto tempo para sentir-se produtivo, mas um começo lento é natural nesse ambiente.

Se receber uma lista de tarefas, deve levá-la a sério, preocupando-se mais com o processo do que com o resultado. As equipes de ciência de dados estabelecidas geralmente têm suas próprias peculiaridades que precisarão ser adotadas. Fazer perguntas nessa fase não é apenas bom, mas, sim, essencial para sua capacidade de realizar o trabalho mais tarde. Os primeiros meses são sua oportunidade de ver o que foi feito antes e aprender sobre o ritmo dos seus colegas.

### **9.1.2 Integração em uma empresa pequena: que integração?**

“Ah, está começando hoje?”. Se estiver ingressando em uma pequena startup, não se surpreenda se nem tudo estiver pronto, incluindo seu computador. Talvez tenha até mesmo que descobrir sozinho como acessar os dados. Pode acontecer que os dados não estejam bem otimizados para seu trabalho e que uma consulta SQL em uma pequena tabela de 100 mil linhas leve seis minutos para ser executada. As sessões de integração para aprender sobre a empresa podem levar semanas para acontecerem, se é que existem, porque não há pessoas suficientes começando em uma determinada semana para fazer sentido ter essas sessões com frequência.

Não há padrões de ciência de dados. Ninguém lhe dirá qual linguagem de programação usar ou como abordar e estruturar uma análise. No entanto, pedirão que comece a obter resultados rapidamente. Ao contrário de uma grande empresa, não precisa se preocupar em não ser produtivo; imediatamente você terá de sê-lo. Mas precisa preocupar-se muito mais se está fazendo acidentalmente algo errado porque ninguém dirá, o que pode levar a descobrir, alguns meses mais tarde, que seu trabalho (incorreto) já está sendo usado. É por isso que ainda é tão importante fazer perguntas e trabalhar com apoio antes de deixar de ser alguém novo na empresa. De uma crise a outra, você se cansará rapidamente; então, trabalhe para criar seus próprios processos que lhe permitirão ser bem-sucedido no longo prazo.



### 9.1.3 Compreender e definir expectativas

Uma das coisas mais importantes que pode fazer nas suas primeiras semanas é reunir-se com seu gerente para discutir prioridades. Essa reunião é importante porque lhe dá conhecimento sobre o que esperar com que trabalhe. Em alguns empregos na área da ciência de dados, a prioridade é fornecer análises a um grupo específico de stakeholders a fim de ajudar a desenvolver uma parte específica da empresa. Em outros trabalhos da ciência de dados, o objetivo é construir modelos de alto desempenho para o site, além daqueles trabalhos nos quais ambos ou nenhum desses objetivos podem ser aplicados.

Talvez sinta que já deveria saber as expectativas do trabalho com a descrição da vaga e com o processo de entrevistas. Embora às vezes seja verdade, muito pode mudar entre o processo da entrevista e o começo do trabalho. Os entrevistadores podem não estar no mesmo período que você ou a organização pode ter mudado antes de você começar. Ao falar com seu gerente o mais cedo possível conseguirá as informações mais atualizadas e terá tempo para discuti-las.

Em teoria, seu gerente tem uma ideia do que você fará, mas está aberto às suas prioridades e pontos fortes. Juntos terão de definir o que significa sucesso em seu trabalho. Em geral, seu êxito está vinculado ao fato de tornar sua equipe e/ou gerente bem-sucedido; se toda a equipe de ciência de dados não estiver trabalhando em direção ao mesmo objetivo, pode ser difícil apoiarem-se uns aos outros. Para definir seu próprio sucesso é preciso entender quais problemas a equipe está tentando resolver e como o desempenho é avaliado. É ajudando a gerar mais receita trabalhando em experimentos para aumentar a conversão ou construindo um modelo de machine learning para auxiliar os agentes de atendimento ao cliente a preverem as preocupações de um cliente, com o objetivo de diminuir o tempo médio gasto por solicitação?

Os objetivos de desempenho geralmente *não* significam “fazer um modelo de machine learning com 99% de precisão” ou “usar o modelo estatístico mais recente em sua análise”. Essas ferramentas ajudam a resolver um problema; não são o próprio objetivo. Se seus modelos e análises lidarem com problemas que as pessoas não se importam, eles são

inúteis. Pensar que o objetivo é desenvolver modelos de maior desempenho é uma concepção errada comum entre as pessoas que ingressam nos seus primeiros empregos na ciência de dados. Faz sentido que essa concepção errada seja comum, pois muitos estudos acadêmicos e cursos educacionais tratam dos vários métodos de executar modelos precisos. No entanto, em última análise, para a maioria dos empregos na ciência de dados, ter modelos altamente precisos não é suficiente para ser bem-sucedido. Coisas como a utilidade do modelo, nível de percepção e capacidade de manutenção são muitas vezes mais importantes. (Os capítulos 10 e 11 discutem essas ideias com mais detalhes.)

Quando se começa em um novo trabalho, não há como saber quais são as expectativas em termos de responsabilidades. Algumas empresas valorizam o trabalho em equipe; pode-se esperar que trabalhe em vários projetos ao mesmo tempo, mas que deixe de lado seu trabalho por um momento para ajudar um colega. Outras empresas pedem que tenha resultados regulares e não há problema em ignorar emails ou mensagens no Slack para concluir seu projeto. A maneira de descobrir se está atendendo às expectativas é reunir-se regularmente com seu supervisor direto. Na maioria das empresas, você terá uma reunião semanal para discutir o que está sendo trabalhando ou outros problemas que houver. Essas reuniões existem para saber se está investindo seu tempo nas tarefas que importam para seu chefe. Por que adivinhar o que esperam se dá para receber feedback explícito? Pensar em blocos de curto prazo ajuda a garantir que está no caminho certo quando as análises de desempenho mais amplas forem feitas.

## **Como se adaptar para obter sucesso**

A menos que esteja em uma empresa muito pequena, haverá um processo formal de avaliação de desempenho; portanto, não deixe de perguntar o que esse processo implica e quando acontecerá. Uma prática comum é ter uma avaliação a cada seis meses, com aumentos salariais e possíveis promoções em seguida. Muitas empresas fazem essa avaliação como um processo 360, no qual recebe feedback direto não apenas do gerente, mas também de seus colegas. Se for esse o caso, descubra se é você ou seu gerente que escolhe os colegas, a fim de entender quem são os stakeholders mais importantes.

As equipes de ciência de dados já consolidadas podem ter uma matriz que mostra em que áreas você é avaliado e o que se espera de cada área em diferentes níveis de antiguidade. Uma área poderia ser a especialização técnica, por exemplo. Um cientista de dados júnior pode ter apenas conhecimento básico e mostrar que está aprendendo; um de nível médio pode ter uma área de especialidade; um cientista de dados sênior pode ser a pessoa a quem a empresa recorre para toda uma área, como testes A/B ou trabalhos de big data. Se uma matriz não existir, veja se pode criar algumas áreas com seu gerente.

Independentemente do sistema, planeje com seu gerente a realização de uma avaliação após os primeiros três meses, se essa prática não for comum. Essa avaliação irá ajudá-lo a certificar-se de que está em sintonia com seu gerente, fornecer atualizações e planejar o resto dos seus primeiros seis meses e ano.

O objetivo de definir o sucesso não é ser excelente em cada área durante os primeiros meses. Na verdade, a maioria das empresas não fará uma avaliação formal de desempenho de alguém que esteja lá há menos de seis meses, porque boa parte desse tempo é investido em aprendizado. Em vez disso, definir o sucesso é certificar-se de aprender a função e começar o trabalho tendo em mente a visão do todo.

### **9.1.4 Como conhecer seus dados**

É preciso aprender também a parte da ciência de dados, claro. Se sua empresa vem fazendo ciência de dados há algum tempo, um ótimo lugar para começar é lendo os relatórios que os funcionários escreveram. Os relatórios mostram não só quais os tipos de dados que sua empresa mantém (e oferecem ideias importantes), mas também o tom e o estilo de como deve comunicar os resultados. Grande parte do trabalho de um cientista de dados é transmitir informações a colegas não técnicos e, ao ler os relatórios, ter uma noção de como são esses colegas não técnicos. Observe como os redatores simplificaram ou deixaram complexos certos conceitos e, assim, será menos provável de explicar pouco ou muito quando chegar o momento de escrever seus próprios relatórios.

Depois, é preciso aprender onde os dados ficam e ter acesso a eles. Ter esse acesso inclui saber qual tabela contém os dados desejados e talvez também qual sistema de dados os possui. É possível que os dados acessados com mais frequência estejam em um banco de dados SQL, mas aqueles referentes a eventos de dois anos atrás estão no HDFS (Hadoop Distributed File System), que precisa de outra linguagem para acessar.

Dê uma olhada geral nos dados com os quais trabalhará regularmente, mas mantenha a mente aberta. Algumas tabelas têm documentação (empacotada com os dados ou em um relatório sobre os dados) que explica problemas potenciais de qualidade ou peculiaridades. Leia esses documentos primeiro, pois eles evitarão a investigação de “mistérios” que, no fim, já foram resolvidos. Depois, dê uma olhada em algumas linhas e estatísticas de resumo. Essas informações podem ajudá-lo a evitar algumas “pegadinhas”, como descobrir que algumas assinaturas iniciam no futuro ou que uma coluna tem muitos valores ausentes. Quando encontrar surpresas

que não estão documentadas, a melhor maneira de entendê-las é conversar com o especialista nessa tabela. Essa pessoa pode ser um cientista de dados, se sua empresa for grande o suficiente, ou alguém que coletou os dados. Talvez descubra que a surpresa seja um problema de verdade que precisa ser resolvido ou talvez seja apenas o esperado. As assinaturas que iniciam no futuro, por exemplo, podem ser aquelas que foram pausadas e definidas para reiniciar naquela data. Ou os cupons para a promoção de Ano Novo referentes ao ano anterior que foram utilizados em maio deste ano podem ter sido usados em maio porque a equipe de suporte os emitiu.

Algumas empresas são melhores do que outras em relação aos dados que foram criados para testes separados de dados reais, embora outras mesquem os dados sem pensar direito. Neste último caso, é melhor perguntar se deve excluir determinados pedidos ou atividades gerados por conta de testes ou parcerias comerciais especiais. Da mesma forma, alguns conjuntos de dados incluem usuários com um comportamento radicalmente diferente. A American Airlines, por exemplo, uma vez ofereceu um passe vitalício que incluía uma tarifa de acompanhante. Uma das pessoas com o passe usou a tarifa para um desconhecido, animal de estimação ou seu violino e pôde voar diversas vezes em um dia. Embora talvez não haja um caso tão extremo, não é incomum para empresas mais novas oferecerem promoções que mais tarde parecem bobas (como 10 anos de acesso por 100 dólares) e podem precisar ser contabilizadas em sua análise.

Ao longo desse processo de investigação dos dados, a questão é entender de forma geral como são seus dados. Se estiver em uma empresa menor, poderá achar que precisa trabalhar com engenheiros para coletar mais dados antes que os dados gerais possam ser úteis. Se estiver em uma empresa maior, decifrárá dezenas de tabelas para ver se existe o que deseja. Talvez esteja procurando tabelas com uma coluna chamada Pedido em 12 bancos de dados. No melhor dos casos, deve haver tabelas bem documentadas e bem mantidas para a métrica principal da empresa, como transações ou assinaturas. Mas é possível que não seja o caso de outros conjuntos de dados menos importantes. Então, tente saber mais se irá se concentrar em uma das áreas menos documentadas.

Aprenda como os dados foram transmitidos para você. Se estiver

trabalhando com algo como dados do site, é provável que será por meio de vários sistemas para ir do site ao banco de dados que pode ser utilizado. Cada um desses sistemas provavelmente altera os dados de alguma forma. Quando a coleta de dados para subitamente, é bom saber onde tentar encontrar o problema (em vez de entrar em pânico). Mas alguns locais têm dados que as pessoas inserem manualmente, como médicos em um hospital ou resultados de pesquisa. Nessas situações, preocupe-se menos com fluxos (pipelines) e muito mais em compreender os vários atributos dos dados e locais potenciais onde um humano inseriu incorretamente. Praticamente em qualquer lugar terá de lidar com alguns dados desorganizados.

À medida que avança, tente anotar quaisquer “pegadinhas” nos dados e fazer um mapa de onde tudo se encontra. É difícil se lembrar desses tipos de fatos no decorrer de um trabalho, e muitas empresas não têm um sistema excelente para a documentação ou descoberta de dados. Assim como colocar comentários no código auxilia você e seus colegas a compreenderem o propósito dele, a documentação de dados fornece enormes dividendos. Embora manter essa documentação localmente no seu computador esteja correta, o melhor a fazer é armazená-la em algum lugar onde todos na empresa possam acessar. Você ajudará novos funcionários e até mesmo cientistas de dados da empresa que não estejam familiarizados com essa área específica.

## **Elin Farnell: opinião sobre a transição da área acadêmica para o mercado de trabalho**

Após oito anos como matemática na área acadêmica, comecei a considerar uma mudança para trabalhar em empresas, quando reconheci que certos aspectos do meu trabalho que mais valorizava eram cruciais a cargos da ciência de dados no mercado de trabalho. Dois dos meus projetos de pesquisa foram colaborações com uma empresa de engenharia, com subsídios do Departamento de Defesa e do Departamento de Energia. Eu amei que, nesses projetos, nosso grupo de pesquisa conseguiu lidar com perguntas matemáticas interessantes, ao mesmo tempo em que sabíamos que o projeto em desenvolvimento seria usado para resolver um problema no mundo real. Também gostei da oportunidade de aprender novos conteúdos matemáticos para resolver o problema e colaborar com uma equipe interdisciplinar. Na minha recente transferência do mundo acadêmico para o mercado de trabalho, alguns aspectos desse novo percurso de carreira destacaram-se pelo contraste com a minha experiência anterior:

- *A amplitude e a profundidade* — na área acadêmica, especialmente para pesquisadores em início de carreira, muitas vezes a prioridade é estabelecer um programa de pesquisa centrado em uma subárea complexa e restrita. Nas empresas, por outro lado, o objetivo é geralmente a resolução de uma grande variedade de problemas, o que significa aprender e utilizar um amplo conjunto de ferramentas da área. Os dois cenários podem ser gratificantes de formas diferentes. A forma como essa amplitude e complexidade se manifestam varia dependendo da instituição e área de pesquisa na universidade ou do foco da equipe e do projeto nas empresas. Suas preferências pessoais quanto a essa questão podem ajudá-lo a avaliar várias oportunidades de emprego.
- *Autonomia* — a área acadêmica oferece uma autonomia significativa em termos de projetos de pesquisa aos quais escolhe se dedicar. Nas empresas, espera-se que os problemas que seu empregador tem sejam solucionados (geralmente com flexibilidade significativa em relação a como você os resolve). Como observei no parágrafo de introdução, a vantagem desse problema é saber se aquilo que está sendo trabalhado terá um impacto positivo no mundo real. Também deve-se observar que existem mecanismos para uma maior autonomia nas empresas;

muitas funções têm a flexibilidade de os cientistas de dados proporem novas áreas para trabalho futuro, e o trabalho de concessão interna ou externa pode disponibilizar tempo e recursos para novos projetos.

- *Equilíbrio entre vida e trabalho* — o entendimento com relação à maioria das pessoas que trabalharam nesses dois mundos e com base na minha experiência pessoal até agora, é que o equilíbrio entre vida e trabalho tende a ser melhor nas empresas. Na universidade, é muito difícil estabelecer limites, e é natural levar trabalho para casa nas noites e nos fins de semana. Embora o trabalho fora do horário regular também seja comum nas empresas, ele é mais orientado a prazos e tende a passar por fases. O equilíbrio entre vida e trabalho é extremamente dependente da cultura da instituição ou de uma empresa em particular e também do modo como você pessoalmente se engaja e contribui para essa cultura. Conheço pessoas nos dois cenários que tiveram dificuldade em gerenciar isso e outras que encontraram um equilíbrio saudável com sucesso.

## 9.2 Como tornar-se produtivo

Em algum momento, você fará seu gerente parecer bom e ter menos trabalho, mas, no início, você dará trabalho a ele, o que é esperado. Leva mais tempo do que se imagina para ser totalmente produtivo. É normal sentir-se frustrado durante esse período, mas lembre-se de que está lidando com uma grande quantidade de carga cognitiva, por estar em um novo ambiente. Você está tentando aprender as normas (provavelmente não ditas) sobre quanto tempo as pessoas levam no almoço, quais são os horários de trabalho, que formas de comunicação usar, se todos desligam o computador quando se afastam da mesa e muito mais. Além disso, ainda tem todo um sistema de dados para conhecer.

É importante salientar que é um erro fácil assumir que é preciso se mostrar competente logo no início, como “preciso fazer tudo mais rapidamente ou, então, eles irão se perguntar por que me contrataram”. É um caso da síndrome do impostor (Capítulo 8). A menos que esteja em uma empresa verdadeiramente disfuncional, eles esperam pela curva de aprendizado. Em vez de querer provar logo seu valor, foque em agregar



valor a longo prazo (em alguns meses, não semanas). No início, você exigirá mais da empresa (“Posso ter acesso a isso? Por que essa consulta é tão lenta?”) do que dar retorno a ela (na forma de relatórios, análises e modelos).

Dito isso, é possível fazer algumas coisas logo no início. Foque em questões simples e totalmente descritivas, como: “Qual é a distribuição do tamanho dos nossos clientes?” ou “Que porcentagem de nossos usuários estão ativos a cada semana?”. No processo, você se familiarizará com os dados da empresa e também encontrará à sua espreita alguns imprevistos e armadilhas. Durante as reuniões com seu gerente, mostre alguns dos seus trabalhos em andamento para saber se está indo na direção certa. É frustrante investir tanto tempo e, no fim, estar respondendo à pergunta errada, usando uma metodologia que seu chefe odeia ou utilizando a fonte de dados errada.

Concentrar-se em questões mais simples também o impede de se constranger, apresentando uma conclusão incorreta por estar tentando responder a uma pergunta complicada sem antes aprender todos os detalhes dos dados. Essa situação pode ser desafiadora, porque se os stakeholders forem novos na ciência de dados, sua primeira pergunta poderá ser uma questão como “É possível prever quais vendas serão concluídas?” ou “Como podemos maximizar a retenção de usuários?”. Mas como falaremos no Capítulo 12, um de seus trabalhos como cientista de dados é aprofundar-se na questão da empresa para encontrar a pergunta de dados subjacente a ela. Se as pessoas não souberem ou tiverem concepções erradas sobre fatos básicos (como a porcentagem de usuários que fazem uma segunda compra ou quantas pessoas clicam em anúncios), eles não farão as perguntas certas.

Duas estratégias podem ajudá-lo a tornar-se produtivo mais rapidamente: fazer perguntas e construir relacionamentos. Fazer perguntas o auxilia a compreender com maior rapidez os detalhes do seu trabalho. Construir relacionamentos permite compreender o contexto da sua função na empresa.

### **9.2.1 Como fazer perguntas**

Uma das coisas que mais podem atrasá-lo em sua carreira é o medo de fazer

perguntas ou dizer “Não sei”. Como já dissemos, a ciência de dados é uma área tão grande que ninguém sabe tudo ou até mesmo 20% dela! Não há como saber todas as complexidades dos dados da sua empresa. Seu gerente com certeza prefere que você faça perguntas e leve alguns minutos do tempo de alguém em vez de ficar tentando reinventar a roda durante dias. Uma pergunta útil pode ser qualquer coisa desde uma pergunta técnica (como “Que teste estatístico usamos para detectar uma mudança na receita em um teste A/B?”) até uma pergunta sobre a empresa (tipo “Qual equipe é responsável por este produto?”).

Dito isso, nem todas as perguntas são criadas da mesma forma. A seguir, algumas sugestões para fazer perguntas melhores:

- *Tente aprender com a observação sobre a cultura de perguntas da empresa.* As pessoas fazem perguntas pessoalmente, em um canal do Slack, em fóruns ou por email? Saber o canal certo significa menos probabilidade de incomodar alguém. Também dá para perguntar ao seu gerente a metapergunta de como fazer perguntas.
- *É bom demonstrar proatividade.* É possível dizer “Pesquisei isto e encontrei estas três coisas” ou “Isto soa como X. É assim?”. Depois de ter feito uma pesquisa sozinho, poderá responder à pergunta por conta própria e poderá questionar com uma melhor ideia do conceito.
- *Não faça perguntas quando puder encontrar as respostas por conta própria rapidamente.* A menos que o tópico seja apresentado enquanto estiver trabalhando com alguém ou discutindo um problema, evite fazer perguntas que sejam respondidas no primeiro resultado do Stack Overflow no Google (como “Qual é a diferença entre um vetor e uma lista na linguagem R?”).
- *Encontre os especialistas e seja cuidadoso com o tempo deles.* Embora algumas de suas perguntas sejam gerais, também é possível haver perguntas técnicas profundas. Descobrir quem são os especialistas em vários métodos estatísticos ou de programação é importante, pois em geral são essas pessoas que darão as respostas. Não seja um fardo para esse tipo de pessoa (ou qualquer outra), então, se perceber que tem muitas perguntas a fazer para uma determinada pessoa, tente agendar

uma reunião com ela. É muito menos provável que as pessoas sintam-se incomodadas se tiverem um horário de reunião em vez de enfrentarem perguntas a cada poucos minutos. Também pode ser bom perguntar o estilo dessa pessoa. Alguns funcionários têm o papel de apoiar outros, mas se essa pessoa também é obrigada a ter resultados, veja se ela tem um calendário que bloqueia determinados momentos em que não está disponível e respeite esses horários.

- *Evite fazer críticas mascaradas de pergunta*, como “Por que você programa assim essa solicitação, em vez da maneira bem melhor que aprendi na universidade?”. Tente compreender por que as coisas são feitas da maneira que são. Se a empresa existe há algum tempo, há muita dívida técnica. Se uma grande empresa tiver servidores físicos, por exemplo, mover esses dados para a nuvem leva mais de meio ano de trabalho para dezenas de engenheiros. Quando perguntam “Por que não fazemos apenas X? É tão fácil e nos pouparia muito tempo”, muitas vezes assumem que outras pessoas não entendem o problema ou sentem que é urgente. Mas a razão pela qual não estão fazendo X pode ser por causa de algo que você não faz ideia, como restrições legais.
- *Acompanhe outra pessoa*. Uma ótima maneira de aprender é acompanhar as pessoas. Em vez de apenas fazer perguntas e obter respostas, é possível ver como as pessoas encontraram essas respostas. Para uma pergunta técnica, acompanhar alguém é também uma forma de ver o seu ambiente de programação e aprender novas técnicas. Mesmo que sua pergunta seja sobre como obter dados, você pode aprender em qual tabela os dados se encontram, como eles sabiam em qual tabela e talvez alguns truques de programação. Seu objetivo final é ter capacidade de responder por conta própria tantas perguntas quanto possível, sabendo onde olhar.
- *Faça uma lista*. Por fim, se tiver perguntas que não precisam de uma resposta imediata, tente manter uma lista de coisas que podem ser úteis para saber, como a frequência com que os dados são atualizados, o limite de tamanho para consultas e a distância dos dados do servidor local. Em seguida, repasse esses itens em um bloco com seu mentor ou gerente. Essa abordagem impede interromper alguém continuamente, o

que pode se tornar um incômodo se não tiver cuidado.

### **9.2.2 Como construir relacionamentos**

Uma parte importante para sentir-se confortável no novo ambiente de trabalho é construir uma rede de apoio. Algumas pessoas constroem mais facilmente do que outras, mas é bom ter conversas que não sejam técnicas com os colegas. Na maioria dos casos, significa fazer reuniões com pessoas com as quais nunca conversou antes a fim de conhecê-las e também saber do trabalho das mesmas. Não é tempo desperdiçado; dessa maneira, você e seus colegas sentem-se mais confortáveis confiando uns nos outros se souberem mais do que apenas nomes e os títulos de cargo.

Aproximar-se de alguém que você não conhece pode ser assustador, mas pode usar suas perguntas como uma forma de iniciar uma conversa. As pessoas gostam de ser úteis e de terem conhecimentos, por isso não tenha medo de usar as perguntas como uma desculpa, desde que seja educado e amigável nas suas dúvidas. Ao conhecer algumas pessoas, até mesmo os escritórios maiores ficam parecendo menos intimidantes. Também é normal que as pessoas mais próximas com quem irá trabalhar agendem uma reunião de 30 minutos para se conhecerem. Se trabalhar em um escritório grande, peça ao seu gerente para fazer uma lista das pessoas que deveria conhecer.

Em um escritório de qualquer tamanho, é bom saber a quem recorrer para perguntas específicas. Uma pessoa pode ser a melhor da empresa em SQL, enquanto outra pode ser responsável pelo sistema de experimentos. É muito útil saber a quem recorrer caso se depare com um obstáculo técnico, e essa pessoa normalmente não é seu gerente. Apresente-se também ao seu chefe de nível mais alto – não para passar por cima do seu chefe, mas porque fica mais fácil quando tiverem que discutir sobre você se essa pessoa já conhecê-lo.

Da mesma forma, tente conhecer todos os stakeholders com quem trabalhará. Se a equipe de ciência de dados for inferior a dez pessoas, tente reunir-se com todas elas individualmente. Se estiver trabalhando com engenheiros de dados ou outras pessoas de dados, fale com eles. Essas reuniões podem ser informais, mas é importante você não existir apenas

como uma assinatura de email. Mesmo que trabalhe sobretudo remotamente, tente usar um sistema de videoconferência para que as pessoas possam ver seu rosto.

Escute bastante, tanto em reuniões oficiais quanto em oportunidades sociais, como um almoço. Conheça as pessoas que trabalham em áreas adjacentes à de dados (o que pode incluir de tudo, desde engenharia até financeiro, operações de vendas e análise de marketing) e escute sobre como conduzem os trabalhos. Não se apresse com uma declaração do tipo “Poderia fazer isso melhor” nem assuma compromissos prematuramente, como “Construiremos uma plataforma de machine learning para isso”. Basta concentrar-se na coleta de informações e pensamentos. E não se esqueça de que nem sempre tudo precisa ser relacionado a trabalho. É bom conhecer as pessoas em âmbito pessoal, seja perguntando sobre os planos para o fim de semana, programas de TV favoritos ou passatempos.

Uma última palavra: seja amigo do diretor do departamento. Os gerentes de escritório controlam muitas das coisas que podem fazer seu dia melhor: lanches, pedidos de almoço, que tipo de sabonete está no banheiro e assim por diante. Eles também têm um dos trabalhos mais difíceis e ingratos – então, faça com que se sintam apreciados.

## **Mentoria e patrocínio**

“Encontre um mentor” é um dos conselhos de carreira mais comuns, mas pode não ser realizável, o que é frustrante. É verdade que ter um *mentor* – alguém que oferece aconselhamento profissional – pode ajudá-lo a resolver questões espinhosas e a tomar melhores decisões. Mas, ao contrário de aprender programação ou de melhorar suas competências de comunicação, o processo de conseguir um mentor não envolve aulas ou livros. Então, como encontrar um?

Felizmente, a mentoria não precisa necessariamente ser de uma relação de longo prazo. Angela Bassa (a quem entrevistamos no Capítulo 16) reuniu uma lista de pessoas dispostas a responder a perguntas e a orientar os recém-chegados da ciência de dados no site [datahelpers.org](http://datahelpers.org). Um mentor não precisa ser alguém a quem chamar a cada dilema de carreira que enfrentar, mas pode encontrar um para auxiliá-lo com um problema específico que está tendo, como praticar entrevistas comportamentais ou fazer seu primeiro pacote em linguagem R.

Determinado tipo de pessoa pode ter ainda mais influência em sua carreira: um patrocinador. Um *patrocinador* é alguém que dá oportunidades às pessoas, seja financiando seu projeto, defendendo sua promoção, apresentando-as a pessoas importantes ou garantindo que elas sejam alocadas para projetos desafiadores que podem ajudá-las a crescer. Mais do que um mentor, é preciso mostrar a um patrocinador que fará um bom trabalho com a oportunidade que estão lhe oferecendo. Se alguém recomenda que você fale em uma conferência, por exemplo, e você não responde ao organizador ou vai para a palestra claramente despreparada, seu comportamento repercute mal na pessoa que o recomendou. Não é preciso ter feito a mesma coisa antes, mas se puder mostrar que fez algo similar (como um meetup), e se for responsivo e educado em suas comunicações com um patrocinador, dá para confiar que você fará bem o trabalho.

Se quiser que alguém seja um mentor ou um patrocinador de longo prazo, mantenha-o atualizado sobre como seguiu seus conselhos ou aproveitou a oportunidade com que o ajudaram. Muitas pessoas são mentores e patrocinadores porque querem ajudar as pessoas, e é gratificante que ouçam como você se beneficiou. E se nunca se comunicou com eles, exceto quando precisou de algo, ele pode sentir

que você está apenas usando-o, o que ninguém quer.

Muitos dos artigos sobre patrocínio e mentoria são sobre encontrar essas pessoas em sua empresa, que é especialmente importante se você trabalhar em uma empresa grande. Mas é comum que os cientistas de dados mudem de empregos a cada poucos anos, e a comunidade de ciência de dados tem poucas oportunidades para que possa começar a construir uma reputação positiva e encontrar patrocinadores e mentores fora da sua empresa que ficarão com você ao longo de vários empregos.

## 9.3 Se você for o primeiro cientista de dados

Até este ponto do capítulo tudo é aplicável aos primeiros meses de qualquer cargo de cientista de dados, mas ser o primeiro cientista de dados em uma empresa tem desafios únicos. Como a área é nova e muitas empresas pequenas não dispõem de cientistas de dados, ser o primeiro não é incomum. Por isso, para ser o primeiro cientista de dados de uma empresa, você deve estar especialmente preparado quando começar.

Quando começar sua nova função, não haverá absolutamente nenhum precedente. Ninguém decidiu ainda se usam Python, R ou alguma outra linguagem de programação. Ninguém descobriu como gerenciar o trabalho. As práticas de desenvolvimento de software, como agile, devem ser usadas para decidir no que trabalhar ou deve-se fazer o que quiser durante dia? Como o código deve ser gerenciado? Deve-se comprar uma licença profissional do GitHub ou um servidor Microsoft TFS usado ou, ainda, dá para manter todos os arquivos na pasta Meus Documentos no computador sem fazer backups?

Como não existem precedentes, tudo o que você fizer se tornará um precedente. Se gostar de fazer seu trabalho na linguagem de programação obscura F#, por exemplo, forçará o próximo cientista de dados a aprender F#. É do seu maior interesse tomar decisões que beneficiarão uma futura equipe, o que pode significar utilizar uma linguagem de programação mais comum do que a sua favorita. Essa abordagem tem de ser equilibrada pelo fato de que focar em excesso no futuro pode causar danos graves ao presente. Se você passar três meses criando um belo pipeline para compartilhar automaticamente relatórios com outros cientistas de dados, mas o segundo cientista de dados não for contratado em cinco anos, esse



trabalho foi um desperdício. Todos os dias, direta ou indiretamente, você tomará decisões que tenham grandes consequências.

Além de ter que descobrir a função por conta própria, você ainda precisa vender a ciência de dados para toda a empresa. Como a empresa não tinha cientistas de dados antes, a maioria das pessoas não entende por que você está ali. Quanto mais rápido as pessoas entenderem sua função, mais provável é que queiram trabalhar com você e manter um cientista de dados por perto. Essas conversas também têm a ver com o gerenciamento de expectativas. Como já discutimos, algumas pessoas pensam que a ciência de dados é basicamente mágica e que o primeiro cientista de dados pode resolver de imediato alguns dos maiores problemas da empresa. É preciso definir expectativas realistas sobre (a) o que a ciência de dados é capaz de fazer e (b) a rapidez com que essas metas podem ser alcançadas. Portanto, seu trabalho exigirá que explique continuamente a ciência de dados em geral às pessoas, bem como o que em particular você pode fazer para auxiliar a empresa. Trabalhar tranquilamente em modelos durante meses é possível se for o 20º cientista de dados de uma equipe, mas certamente não fará isso sendo o primeiro.

Embora ser o primeiro cientista de dados dê muito mais trabalho e seja muito mais arriscado do que outros cargos, também há grandes recompensas. Ao tomar decisões técnicas, você pode escolher coisas que estão mais alinhados com o que quer. Ao vender a ciência de dados à empresa, você começa a se tornar mais conhecido e mais influente. E à medida que a equipe de ciência de dados cresce, mais próximo você fica de liderá-la, o que pode ser excelente para o crescimento da sua carreira.

## **9.4 Quando o trabalho não é o que foi prometido**

Pode ser frustrante começar um trabalho de ciência de dados e descobrir que não está perto daquilo que esperava. Depois de meses de trabalho, finalmente você começou na área e agora talvez tenha que sair e começar tudo de novo. Pior ainda é sair rapidamente, o que não será nada bom para seu currículo, causando-lhe preocupação. Então, significa ter que aguentar um ano? Administrar um ambiente ruim e decidir se deve sair ou não é

desafiador. Nas próximas seções, abordaremos duas categorias principais de problemas: o trabalho é assustador e o ambiente de trabalho, tóxico. Embora não haja uma solução mágica para resolver esses problemas, trataremos de algumas estratégias de mitigação.

### **9.4.1 O trabalho é terrível**

Primeiramente, analise profundamente suas expectativas. O problema é algo que está na linha de: “Todos os meus dados não estão limpos! Passei dois dias apenas na preparação dos dados! Os engenheiros de dados não corrigem as coisas imediatamente!”. Esses problemas farão parte de todas as funções da ciência de dados. Até mesmo cientistas de dados das maiores empresas que contam com centenas de engenheiros têm esses problemas; há tantos dados que é impossível que tudo seja perfeitamente verificado. Embora as tabelas principais devam ser limpas e bem documentadas, provavelmente encontrará dados em subáreas que precisam melhorar ou trabalhar com outras pessoas para coletar.

Uma maneira de verificar o quão realistas são suas expectativas é confirmá-las com outros cientistas de dados. Se você tiver se formado em uma área relacionada ou feito um bootcamp, pergunte a seus colegas ou pessoas na rede de antigos alunos o que pensam do ambiente de dados em que está trabalhando. Se ainda não conhece muitos cientistas de dados, tente ir a meetups ou participar de comunidades online caso esteja em uma cidade pequena ou na área rural. (Tratamos desse tópico em maior profundidade no Capítulo 5.) Se outros cientistas de dados da sua empresa tiveram empregos anteriores em ciências de dados, veja se podem comparar os ambientes.

Outra situação é que o trabalho pode ser entediante e chato. A atividade para a qual foi contratado pode ter sido fazer previsões, por exemplo, mas, na prática, tudo o que faz é pressionar o botão de repetição no modelo de previsão existente de outra pessoa uma vez por mês. Nesse caso, veja se consegue arrumar alguns projetos paralelos na empresa ou automatizar alguns processos. Se o trabalho é chato, mas não consome muito tempo, aproveite a oportunidade para fazer coisas relacionadas à ciência de dados. Continue a montar seu portfólio de ciência de dados com projetos paralelos,

escreva posts de blog ou faça cursos online. Essas táticas o ajudarão na sua próxima função.

Por fim, dá para aprender mesmo com trabalhos ruins. Há alguma maneira de ajustar seu trabalho de modo que faça mais coisas para aprender? Em que áreas pode melhorar? Talvez seus colegas não o ajudem a aprender a escrever um código melhor, mas dá para aprender sobre quais erros são fáceis de cometer na criação de uma equipe de ciência de dados? É muito provável que existam pessoas inteligentes e bem-intencionadas na sua empresa, por isso o que aconteceu para tornar o trabalho desagradável? Ao saber do que não gosta, saberá, também, o que deve considerar na sua próxima busca de emprego e estará mais preparado para evitar erros ao começar sua própria equipe de ciência de dados.

### **9.4.2 O ambiente de trabalho é tóxico**

A Seção 9.4.1 discute uma situação ruim, mas gerenciável. Mas e se seu trabalho for realmente tóxico? E se seu gerente e stakeholders tiverem expectativas completamente irrealistas e ameaçarem demiti-lo porque não está progredindo na previsão do valor vitalício quando não têm dados para isso? Ou é penalizado quando suas respostas não atendem às expectativas da empresa? As empresas que são novas em ciência de dados podem esperar que faça a mágica da ciência de dados para resolver os seus principais problemas. Podem falar: “Construir um modelo para dizer se o texto está bem escrito” – um problema que ninguém na área resolveu. Nesse caso, é preciso ajustar as expectativas da empresa ou arriscar a se sentir como tendo um desempenho ruim constantemente. Defender-se nessa circunstância é difícil, mas em geral há pessoas sensatas e atentas trabalhando em qualquer empresa. Se disserem: “Se fosse um cientista de dados melhor, seria capaz de fazer isso”, trata-se de um grande sinal de alerta. Mesmo que um cientista de dados muito mais experiente pudesse lidar com o problema, a empresa deveria ter reconhecido esse fato quando estava projetando e contratando para a função.

Talvez o problema seja que as pessoas e as equipes não estão colaborando. Em vez de observarem onde podem ajudar, as equipes podem estar tentando sabotar umas às outras. Elas se concentram apenas em como

chegar à frente e podem até ver que poderão obter sucesso na empresa como um jogo de soma zero: se você ou sua equipe está indo bem, a deles está perdendo. Além de criar um ambiente nada saudável, essa situação geralmente demanda muito trabalho desperdiçado, já que pode acabar duplicando o projeto de outra pessoa porque não compartilhariam os dados nem aprenderiam com você.

O problema pode não ter nada a ver com a parte da ciência de dados; o ambiente pode ser sexista, racista, homofóbico ou de outra forma hostil. Não há por que sentir-se desconfortável ao ir trabalhar todos os dias. Mesmo que não seja abertamente hostil, ser interrompido em reuniões, ser tratado por pronomes incorretos ou ser questionado: “Mas de onde *realmente* você é?”, pode concorrer para que se sinta inadequado.

Infelizmente, resolver esses tipos de problemas requer normalmente o envolvimento da liderança e compromisso ativo de todo o mundo. Mas o fato de o local de trabalho ser tóxico indica muitas vezes que a liderança inexistente ou até mesmo está contribuindo ativamente para o problema. Se o problema está enraizado em uma má pessoa, idealmente outros reconhecerão esse fato, e a pessoa será removida. Mas se o problema for generalizado, pode ser impossível mudar, e tentar mudá-lo sendo um funcionário júnior é uma receita rápida para ficar cansado. Nessas situações, é melhor pensar com cuidado em se é preciso ir embora.

### **9.4.3 Decidir ir embora**

Decidir se é melhor ir embora consiste em uma decisão extremamente pessoal. Embora ninguém possa lhe dar um fluxograma simples que fará a decisão ser indolor e fácil, podemos oferecer algumas perguntas para pensar e que podem orientar sua decisão:

- Você tem economias suficientes, um cônjuge com outra renda que possa apoiá-lo ou familiares a quem pedir um empréstimo se for embora sem ter outro emprego?
- Seu trabalho está afetando sua saúde ou vida fora dele?
- Se o problema for o trabalho, chegou a conversar com o gerente sobre os problemas e tentou resolvê-los?

- É possível mudar de equipe ou função – se não agora, dentro de alguns meses?

Se as respostas a essas perguntas o fazem sentir que é melhor ir embora, uma opção é começar a procurar imediatamente outros empregos. Mas pode ser que se preocupe em ter a marca de um trabalho de curto período no seu currículo ou em como explicar isso aos entrevistadores. Se estiver no trabalho há apenas algumas semanas e tiver chegado diretamente do último emprego, considere entrar em contato com seu gerente anterior. É provável que seu cargo ainda não tenha sido preenchido e, se tiver saído bem de lá, poderá voltar.

Se estiver buscando um novo emprego, a seguir apresentam-se algumas dicas de como falar sobre essa curta experiência na entrevista:

- *Aguarde que o entrevistador toque no assunto.* Não sinta que precisa falar sobre isso de forma proativa; pode não ser uma preocupação, em especial porque a empresa está claramente interessada. (Chegou ao estágio da entrevista!)
- *Encontre alguma experiência positiva e aprendizado sobre o trabalho para falar.* Essas experiências e lições podem ser um projeto em que trabalhou, exposição ao setor ou orientação de um líder sênior.
- *Quando perguntarem por que razão saiu tão logo, dê uma explicação breve e neutra.* Você se encontra em uma situação difícil porque quer ser honesto sobre a razão pela qual saiu e que não foi por falha sua, mas, se for aberto e honesto demais, o entrevistador pode injustamente percebê-lo como alguém difícil de se trabalhar. Assim, sua melhor aposta é falar algo vago, como: “Os requisitos do meu trabalho não eram o que eu esperava, e não conseguia usar minhas competências e experiência para beneficiar a empresa” e deixe assim. Se aprendeu algo sobre o tipo de ambiente de trabalho desejado, fale sobre isso. Talvez fosse o primeiro cientista de dados em uma empresa e percebeu que não quer integrar uma equipe maior.

Se decidiu ir embora, consulte o Capítulo 15 para obter informações sobre como fazê-lo com jeito, incluindo buscar um novo emprego enquanto trabalha em tempo integral e sair bem da empresa.

Mas talvez não possa sair porque seu visto está ligado ao local de trabalho ou a empresa é a única que faz ciência de dados em sua pequena cidade. Se for essa a sua situação, veja algumas dicas a seguir:

- *Lembre-se de que você não é seu trabalho.* Você não tem que assumir a responsabilidade pelas más decisões que a empresa toma. A menos que esteja em uma posição de liderança, provavelmente terá pouco controle sobre o que a empresa faz.
- *Tente se manter saudável.* Não sacrifique sono, exercícios e tempo com amigos e familiares.
- *Fale com alguém.* Talvez essa pessoa seja um cônjuge, um amigo ou um terapeuta. Eles podem dar conselhos, mas apenas ouvir o que você tem a dizer já o ajudará.
- *Pense em reportar assédio pessoal.* Se estiver sendo assediado por uma pessoa específica, considere reportar ao departamento de recursos humanos. Documente tudo. Não apareça nos recursos humanos pessoalmente; envie emails e receba emails de volta para que tenha um registro do processo. Talvez tenha que apontar determinadas coisas que lhe foram ditas, e ter isso registrado por escrito será útil. Se a empresa não fizer nada e se estiver nos Estados Unidos, poderá apresentar uma reclamação na Comissão para a Igualdade de Oportunidades de Emprego. Infelizmente, denunciar assédio tem seus riscos: embora seja ilegal, há empresas que retaliaram os funcionários pela denúncia e que atrapalharam seu crescimento profissional ou até mesmo os demitiram. Mesmo que não queira denunciar o assédio, considere manter a documentação de qualquer assédio caso decida denunciar mais tarde.
- *Pense se há alguma outra ideia além de ir embora da empresa.* Talvez sinta que não pode ir embora porque as únicas opções disponíveis é ir para uma empresa menos prestigiada, ter um cargo mais baixo ou esgotar temporariamente suas economias. Mas não subestime os efeitos negativos de permanecer em um ambiente tóxico: se puder fazer um sacrifício a curto prazo para ir embora, provavelmente valerá a pena no longo prazo.

Esperamos que nunca se encontre nesse tipo de situação, mas é útil ter um

pouco de informação para “quebrar o vidro em caso de emergência” em algum lugar. Lembre-se de que mudar de emprego na área da ciência de dados é comum (como discutiremos mais adiante no Capítulo 15), por isso não há razão para continuar em um local de trabalho que o deixe desconfortável.

## **9.5 Entrevista com Jarvis Miller, cientista de dados do Spotify**

Jarvis Miller formou-se em 2018 com mestrado em estatística e trabalha como cientista de dados na Missão de Personalização no Spotify, concentrando-se em melhorar a experiência de audição de cada usuário. Quando essa entrevista foi conduzida, ele estava trabalhando como cientista de dados no BuzzFeed.

### **O que o surpreendeu em seu primeiro trabalho de ciência de dados?**

Duas coisas que me surpreenderam foram o quanto eu poderia melhorar como escritor e como precisava explicar minha contribuição de ciência de dados para a empresa sem usar jargão. Eu tinha essa ideia de que, como os stakeholders tinham trabalhado com cientistas de dados, eles haviam aprendido a entender a linguagem e, portanto, eu não precisava mudar a maneira como explicava as coisas. Percebi que não era o caso, e eu não podia simplesmente dizer: “Executei uma regressão logística nesses dados para classificá-los...”. Quanto a ser um escritor melhor, comecei a dar forma à história quando escrevo um relatório; melhorar minhas capacidades de narração de dados e explicar as coisas de uma forma para que gerentes de produtos, designers e stakeholders que não são da área de tecnologia conseguissem entender o que eu estava dizendo.

Vim da área acadêmica, onde sentia que tudo se tratava sobre encontrar ou não o resultado no fim do dia; não importava se começou a trabalhar com o prazo quase vencendo ou se planejou com antecedência. Nas empresas, há um grande objetivo geral, mas descobre como dividi-los em versões. Você faz com que a primeira versão funcione, você a envia, aprende se está indo

bem ou não e talvez a melhor em um trimestre futuro. Eu estava acostumado a seguir até a conclusão. Mas aqui tive de aprender a como priorizar partes de um projeto e finalizá-las. Aprendi a documentar o que tinha feito e o que fazer na próxima versão e depois torná-la compartilhável, seja colocando um relatório em uma pasta compartilhada ou fazendo uma aplicação para que as pessoas possam utilizar meu trabalho e ver o que é para ser feito.

### **Quais são alguns dos problemas enfrentados?**

Dar opiniões foi algo com que tive dificuldade. Quando comecei, estava fazendo um projeto isolado, e a pessoa a quem eu me reportava estava em Nova Iorque, enquanto eu estava em Los Angeles. Se eu ficava confuso, não sabia se deveria enviar uma mensagem imediatamente ou guardá-la para nossa reunião. Sabia que não queria falhar por algo que bloqueava meu trabalho, mas sequer tinha certeza quando algo estava me bloqueando. Penso que se trata de um problema comum para os cientistas de dados, especialmente para aqueles que se encontram em grupos marginalizados ou que vieram de uma área diferente. Eles podem se sentir como se fossem novos ou que não são especialistas e podem não expressar desagrados ou uma opinião. Se pudesse voltar no tempo, teria uma conversa mais cedo sobre como me sentia isolado e não tinha certeza de como a comunicação funcionava nessa empresa.

### **Pode nos contar sobre um de seus primeiros projetos?**

Um deles foi renovar nossa plataforma de testes A/B, que era um problema muito amplo. Comecei pegando uma lista de pessoas para falar sobre o que fizeram no BuzzFeed, como trabalhavam e como os testes A/B encaixavam-se nesse fluxo de trabalho. Depois, discutimos a ferramenta específica: o que eles não gostavam e por que e qual foi o fluxo de trabalho ao usá-la? Infelizmente, levou à questão de assumir muita coisa. Inúmeras pessoas tinham várias sugestões, dei a todas o peso igual, que acabaram virando 50 coisas grandes que precisava fazer. Mas meu gerente me pediu para dividir essas sugestões em coisas totalmente necessárias e boas para se ter, incluindo as razões pelas quais foram priorizadas dessa forma. Ele sugeriu



listar o objetivo geral do projeto e dar pesos a ideias com base na contribuição para o objetivo, assim como quanto tempo levariam.

## **Qual seria seu melhor conselho para os primeiros meses?**

Lembre-se de que foi contratado por um motivo: eles respeitam seu ponto de vista e pensam que podem ajudá-lo a aprender e também a aprender com você. Se tiver uma opinião, tente falar com alguém. Se odeia falar em um grupo grande, talvez mande mensagem a uma pessoa, fale com ela e repense bem antes de expressar essa ideia na frente de um grupo maior.

Isto não se aplica apenas à parte técnica do trabalho. Nos primeiros minutos da reunião com meu gerente, não quero relatar de imediato o que fiz. Início com uma conversa casual por alguns minutos para desestressar e clarear a mente. Sei que ajuda na minha produtividade e que a empresa quer que eu seja produtivo, por isso a informo. Sua opinião é valiosa e vale a pena compartilhá-la, especialmente se for sobre como gosta de ser tratado ou de como pode ser mais produtivo e crescer na função, porque eles não o conhecem como você mesmo e ser produtivo beneficiará a todos.

## **Resumo**

- Não se preocupe em tornar-se totalmente produtivo de imediato. Em vez disso, concentre-se em construir relacionamentos, ferramentas e compreender os dados, o que fará com que você seja produtivo a longo prazo.
- Se estiver em uma situação ruim de trabalho, tente trabalhar para obter controle e mitigar o impacto na sua saúde e carreira.

# CAPÍTULO 10

## Como fazer uma análise eficaz

Este capítulo abrange:

- Planejamento de uma análise
- Trabalho com código, dados e estrutura de projeto
- Entrega da análise ao cliente

Este capítulo foi escrito no contexto de cientistas de dados que se concentram na ciência e análise de decisões – profissionais que usam dados para fornecer ideias e sugestões para a empresa. Embora os engenheiros de machine learning também tenham de fazer análises antes de construir e implementar modelos, alguns dos conteúdos em torno da gestão dos stakeholders com visualizações bonitas são menos relevantes. Se for engenheiro de machine learning e estiver lendo este livro, não se preocupe; este capítulo ainda é muito relevante e irá gostar ainda mais do Capítulo 11, o qual trata da implantação de modelos na produção.

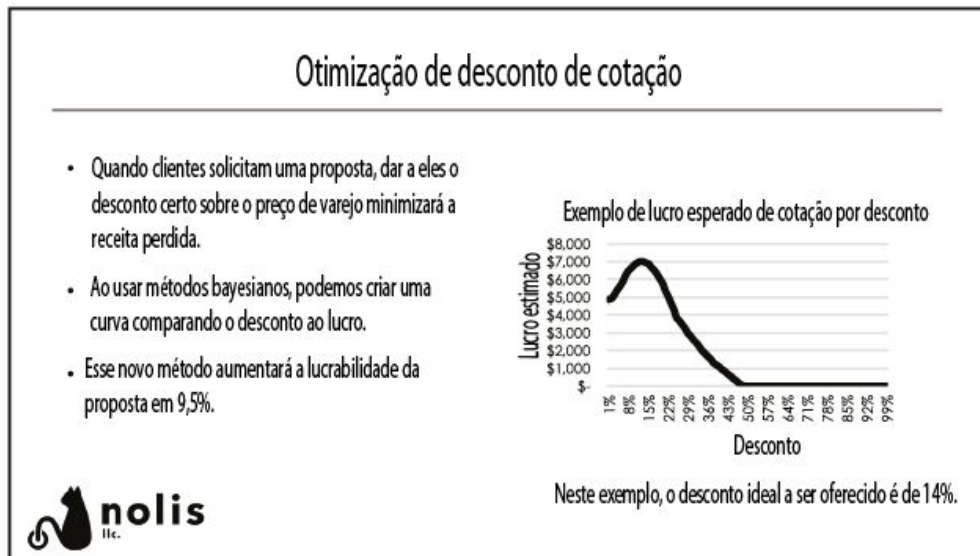
O pilar de muitos trabalhos da ciência de dados está em fazer análises: documentos breves que usam dados para tentar explicar uma situação da empresa ou resolver um problema da empresa. As empresas modernas são construídas com base em relatórios e análises. As pessoas que tomam decisões não estão à vontade para fazê-lo sem dados para sustentar suas escolhas, enquanto os cientistas de dados são algumas das melhores pessoas para encontrar significado nos dados. As análises também são importantes para a construção de ferramentas de machine learning, pois, antes de se construir um modelo de machine learning, é necessário compreender o contexto do conjunto de dados. Criar uma análise que possa pegar a grande quantidade de dados da empresa e convertê-los em um resultado conciso que esclareça o assunto em questão é extremamente difícil e praticamente uma arte. Como esperar que uma pessoa pegue tabelas com milhões de

registros sobre informações históricas, cada uma com complexidades e nuances, e as transforme em um: “Sim, os dados dizem que essa ideia é boa”? O ato de descobrir o que é significativo matematicamente, com o que a empresa se preocupa e como preencher a lacuna entre os dois não é algo que você deveria saber fazer com naturalidade.

Neste capítulo, vamos analisar os princípios básicos de como construir uma análise para entender como fornecer análises significativas à empresa. Usando as competências vistas no capítulo, você será capaz de progredir mais rapidamente na sua carreira em ciência de dados.

O que é de fato uma análise? Uma análise é normalmente um PowerPoint, um arquivo em PDF ou Word ou uma tabela em Excel contendo informações sobre dados e visualizações dos mesmos que podem ser compartilhadas com cientistas que não sejam de dados. A Figura 10.1 é um exemplo de um slide que poderia ser encontrado em uma análise. Uma análise geralmente leva de uma a quatro semanas para ser feita, com um cientista de dados coletando dados, executando código para métodos estatísticos e gerando o resultado final. Quando concluído, o código não é tocado até que a análise seja reexecutada meses mais tarde, ou possivelmente nunca mais. Exemplos de análises incluem:

- Analisar os dados de pesquisa do cliente para ver quais produtos geram maior satisfação.
- Analisar os dados das localizações de onde os pedidos estão sendo feitos para escolher o local de uma nova fábrica.
- Usar dados históricos do setor aéreo para prever quais cidades precisarão de mais rotas.



*Figura 10.1 – Um exemplo de slide de uma análise em PowerPoint.*

Esses exemplos têm níveis variáveis de complexidade técnica; alguns requerem apenas a sumarização e a visualização de dados, enquanto outros necessitam de métodos de otimização ou de modelos de machine learning, mas todos respondem a uma única pergunta.

## **Criação de relatórios x realização de uma análise**

Um relatório e uma análise são similares, mas não são iguais. Um *relatório* é algo gerado de forma recorrente sem muita alteração estrutural entre as versões. O relatório financeiro mensal, por exemplo, pode ser uma grande tabela em Excel que é atualizada com novos números todos os meses. O objetivo de um relatório é manter as pessoas cientes de como as métricas estão mudando. Uma análise é algo feito uma só vez para responder a uma pergunta mais complexa. Uma análise da aquisição de clientes pode ser feita em linguagem R sobre como novos clientes estão comprando produtos, com os resultados colocados em uma apresentação em PowerPoint. Os relatórios tendem a ser preenchidos com números e métricas, enquanto as análises concentram-se em fornecer um único resultado principal. A maioria das características de uma boa análise serve para um bom relatório, portanto, neste capítulo, usamos a análise para falar dos dois, a menos que explicitamente indicado de outra forma.

Então, o que é uma boa análise? Uma boa análise tem cinco características seguintes:

- *Responde à pergunta.* Uma análise começa com alguém fazendo uma pergunta, portanto, para que a análise seja significativa, ela necessita ter uma resposta. Se a pergunta apresentada fosse: “Qual destes dois sites faz com que mais clientes comprem produtos?”, a análise deve mostrar qual site gera mais vendas. Essa resposta pode até ser “não temos informações suficientes para dizer”, mas deve ser uma resposta direta à pergunta.
- *É feita rapidamente.* As respostas às perguntas da empresa influenciarão decisões que têm prazos. Se a análise demorar tempo demais para ser gerada, a decisão será tomada sem a análise. Uma expectativa comum é que a análise seja concluída dentro de um mês.
- *Pode ser compartilhada.* A análise precisa ser compartilhada não apenas com a pessoa que solicitou a realização da análise, mas também com quem ela quiser compartilhar. Se a análise envolver um gráfico, por exemplo, ele não pode simplesmente estar em um script em R ou Python; tem de estar em um formato que as pessoas podem digerir, como o PowerPoint.

- *É autocontida.* Como não se pode prever quem verá a análise, ela precisa ser compreensível por si só. Gráficos e tabelas devem ter descrições claras, os eixos devem ser nomeados, as explicações na análise devem ser escritas e a análise deve evitar, se possível, referenciar outros trabalhos.
- *Pode ser revisitada.* A maioria das questões será feita novamente no futuro. Às vezes, responder a elas significa refazer exatamente o mesmo trabalho, como reexecutar um agrupamento de dados (clustering). Outras vezes, você tem que usar a abordagem em algum outro lugar, como mudar os dados da inserção de clientes europeus para clientes asiáticos.

Essas características agregam-se ao tema geral “uma boa análise é algo que ajuda os cientistas que não sejam de dados a fazerem o trabalho deles”.

O restante deste capítulo está estruturado para tratar cronologicamente das etapas de uma análise, começando com a solicitação inicial de análise e terminando com a apresentação dos relatórios. Embora nem todas as análises sigam esses passos, a maioria segue (ou deveria seguir). À medida que se torna mais familiarizado com a realização de análises, pode sentir-se inclinado a ignorar alguns passos, mas esses atalhos são as mesmas ações que fazem com que cientistas de dados sênior cometam erros.

## **Análises para diferentes tipos de cientistas de dados**

Dependendo da sua função como cientista de dados, as situações em que fará análises variam muito:

- *Cientista de decisão* – para esses tipos de cientistas de dados, fazer análises é a função central do trabalho. Os cientistas de decisão estão continuamente se aprofundando nos dados para responder às perguntas, as quais precisam ser comunicadas à empresa. Uma análise é a ferramenta-chave para isso.
- *Engenheiro de machine learning* – embora um engenheiro de machine learning concentre-se na criação e implantação de modelos, as análises ainda são ferramentas úteis para compartilhar o desempenho dos modelos. As análises são utilizadas para mostrar o valor na construção de um novo modelo ou como os modelos mudam ao longo do tempo.
- *Analista* – são cientistas de dados que se concentram fortemente em métricas e KPIs para o negócio, geralmente se encontram fazendo muitos relatórios. Eles criam um fluxo de dados recorrentes para a empresa, muitas vezes em Excel, SQL, R ou Python. Embora esses especialistas em analítica façam análises, eles precisam pensar sobre a capacidade de manutenção do trabalho mais do que as outras funções, pois eles têm de repeti-lo com muita frequência.

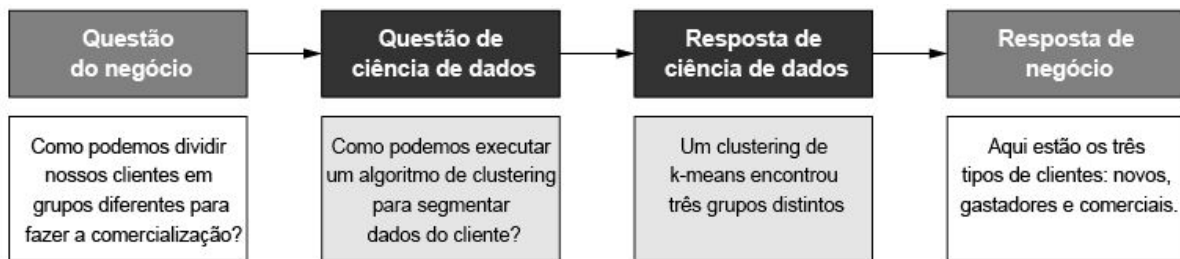
## 10.1 Solicitação

Uma análise começa com uma solicitação de resposta a uma questão de negócios. Alguém de um setor diferente da empresa ou seu gerente fará uma pergunta como: “Pode analisar por que as vendas de widgets foram baixas na Europa em dezembro?” ou “Nossos clientes de empresas pequenas têm um comportamento diferente dos nossos clientes maiores?”. Dependendo do nível de especialização técnica da pessoa que questiona, você pode obter uma solicitação mal feita (“Por que as vendas estão em baixa?”) ou uma bem elaborada (“Quais atributos estão correlacionados a um valor de pedido médio inferior?”).

A análise é formada em torno da questão de negócios, mas não dá para fazer ciência de dados em uma questão de negócios. As questões relacionadas à ciência de dados são do tipo: “Como é que se agrupam esses

pontos de dados?” e “Como fazemos previsões de vendas?”. O cientista de dados tem de fazer o trabalho de converter essa questão de negócios em uma questão de ciência dos dados, respondendo à questão de ciência dos dados e dando uma resposta para a empresa. Esse trabalho é complicado. Compreender como as questões de ciência de dados e as de negócios se relacionam exige uma combinação de experiência com o tipo de problema em mãos e uma compreensão de como os resultados de diferentes métodos estatísticos poderiam ser úteis. Esse fluxo de trabalho de questões de negócios para perguntas e respostas de ciência de dados e, finalmente, de volta para uma resposta para a empresa foi idealizado por Renee Teate, que também é entrevistado no Capítulo 14 deste livro.

A Figura 10.2 mostra graficamente esse processo. Questões de negócios vêm de stakeholders, que querem saber como direcionar o marketing a distintos clientes. O cientista de dados tem de descobrir o que essa solicitação significa em termos matemáticos – nesse exemplo, um agrupamento de dados (clustering) do cliente.



*Figura 10.2 – O processo de resposta a uma questão de negócios com ciência de dados, idealizado por Renee Teate.*

Quando o processo é concluído, o cientista de dados tem uma resposta de ciência de dados (como um conjunto de três grupos de pontos de dados agrupados (clustered)). Por fim, o cientista de dados tem de converter essa resposta em algo que a empresa entenderia, por exemplo grupos como “novos clientes” ou “pessoas que gastam bastante”.

Antes de começar a examinar os dados e a escrever um código para resolver uma questão de ciência de dados, tem de fazer o trabalho básico para compreender melhor a questão de negócios. É preciso entender qual é o contexto para a análise para, então, oferecer algo útil. Quem pediu que a análise fosse feita e qual é sua relação com a equipe dessa pessoa? Qual é o



motivo? Eles têm uma questão muito específica que querem que seja respondida ou uma ideia geral e vaga de um problema, com a esperança de que os dados possam ser úteis? Parece que há dados para resolver esse problema? Se não, o que seria necessário para obtê-los? Fazer perguntas não só ajuda a entender a solução do problema, mas também ajuda a entender para que ele será usado. Muitos cientistas de dados passam semanas em análises apenas para descobrir que não havia necessidade, porque os stakeholders estavam sendo “apenas curiosos”.

Essas questões são geralmente respondidas em uma reunião de 30 a 60 minutos com a pessoa que fez a solicitação, além de mais alguém envolvido no trabalho. Como é você que faz a análise, talvez não seja a pessoa a organizar uma reunião, mas, se não houver ninguém em seu calendário, é melhor agendar uma. Se não foi encontrado antes o solicitante da análise, a reunião é um bom momento para se apresentar e aprender sobre o trabalho que ele desenvolve.

Um conjunto hipotético de exemplos de conhecimento básico seria algo assim:

- *Quem está solicitando a análise?* Júlia da equipe de produtos de widget a solicitou.
- *Qual é o motivo?* As vendas de widgets caíram 10% neste mês, e a equipe de negócios não sabe por quê.
- *Qual é a solicitação?* A equipe quer usar dados para ver se a vendas de widgets foi focada em uma região do país.
- *Qual decisão será tomada?* A decisão é se o produto de widget deve ser descontinuado ou não.
- *Temos os dados necessários?* Sim, a análise precisa dos pedidos de clientes pelo código postal de envio, que está disponível na base de dados de pedidos.

Saber se há dados que possam responder de forma plausível à questão é *realmente* importante. A última coisa que alguém iria querer fazer é gastar várias semanas trabalhando em uma análise e não ter o que dar de retorno ao stakeholder com algo que poderia ser usado.

Um exemplo de uma situação em que não haveria dados seria algo como

o seguinte: em uma empresa de varejo, o stakeholder quer saber quantos pedidos cada cliente fez, mas como os clientes pagam em dinheiro, não há forma de utilizar os dados existentes para dizer quem fez cada pedido. Nesse tipo de situação, é melhor ser honesto com todas as pessoas envolvidas e informá-las que o pedido não é possível de atender. Outras pessoas podem propor formas alternativas de usar dados que possam estar suficientemente próximos do que esperava ou pode ter que explicar porque as alternativas também não funcionarão. Se possível, um dia proponha um plano que possa obter os dados necessários. No exemplo anterior, um programa de fidelidade permitiria que os pedidos fossem associados a um determinado cliente e, assim, resolveria o problema dos dados, embora esse programa levasse tempo para ser criado.

As outras questões, como quem é a pessoa e por que ela está fazendo a solicitação, são úteis para a criação do plano de análise.

## 10.2 Plano de análise

Para os cientistas de dados, nada é mais divertido do que mergulhar diretamente em alguns dados para responder a questões. Vamos carregar dados! Agrupe! Sintetize! Configure um modelo e gere resultados! Infelizmente, com um número infinito de maneiras de resumir e modelar dados, é possível passar semanas trabalhando com dados só para descobrir que nada do que você tenha produzido responde à questão proposta pela empresa. A constatação de não ter feito algo relevante é o *pior*. E acontece com frequência com cientistas de dados, especialmente aqueles em início de carreira.

Uma solução para esse problema é ter uma meta para não se desviar do caminho e fazer um trabalho relevante. Um plano de análise é esse roteiro. A ideia é que, antes de você começar a examinar os dados, escreva tudo o que pretende fazer com eles. Depois, à medida que sua análise avança, mantenha um controle do quanto do plano foi concluído. Quando você tiver finalizado tudo no plano, está pronto! Não só é uma maneira de saber se você está fora do plano, mas também ter uma ferramenta para acompanhar o progresso e manter-se responsável. Pode até utilizá-lo em reuniões com

seu gerente para discutir como as coisas vão indo.

Ao fazer um plano de análise, é melhor que o trabalho seja acionável. “Fazer uma regressão linear nas vendas por região” é algo que pode ser escrito em código, já “descobrir por que as vendas diminuíram?” não é algo direto de se fazer, mas o resultado de realizar outras coisas. Se as tarefas do plano forem pontos de ação, será fácil saber se está progredindo. Também facilitará a análise, pois não terá de se preocupar com o que fazer a seguir. Em vez disso, poderá examinar o plano de análise e selecionar a próxima tarefa a ser executada.

Para fazer seus primeiros planos de análise, recomendamos incisivamente o seguinte modelo:

- *Topo* – liste o título da análise, quem você é (no caso de a análise ser compartilhada com outros) e o objetivo da análise.
- *Seções* – cada seção deve ser um tópico geral na análise. O trabalho de análise realizado em cada seção deve ser autocontido (não contando com o trabalho de outras seções); portanto, deve ser possível que uma pessoa diferente faça cada seção, e cada seção deve ter uma lista de tarefas.
- *Primeiro nível das listas de seção* – o primeiro nível das listas de seção deve ser cada questão que foi feita. Essa seção ajuda todos a lembrarem o porquê desse trabalho específico e, se todas as questões forem respondidas com sucesso, o tópico da seção principal deve ser considerado como compreendido.
- *Segundo nível das listas de seção* – o segundo nível das listas deve ter as tarefas reais a serem feitas, as quais podem ser marcadas como trabalho sendo realizado. Essas tarefas podem ser tipos de modelos a serem executados, por exemplo, e as descrições devem ser específicas o suficiente para que, a qualquer momento, você possa dizer se o trabalho foi efetivamente concluído.

A Figura 10.3 mostra um exemplo de plano de análise, no caso para avaliar por que os clientes estão saindo da região América do Norte. No topo estão o título, o objetivo e as informações de contato do cientista de dados, caso o material seja repassado. Cada seção do plano abrange um componente

diferente da análise (por exemplo, analisar na América do Norte ou comparar com outras regiões). As subseções (numeradas) são questões na análise, enquanto a seção inferior (com letras) é a tarefa específica a ser feita.

Ao criar seu plano de análise, compartilhe-o com seu gerente e com o stakeholder que fez a solicitação. Eles devem dar sugestões sobre como melhorá-lo ou aprovar o trabalho. Um plano de análise aprovado fornece uma base acordada para o trabalho. Se, depois de fazer a análise, o stakeholder perguntar por que as coisas foram feitas dessa maneira, dá para se referir ao plano de análise e aos objetivos originais.

É provável que, à medida que faz a análise, perceba que deixou algo importante de fora do plano da análise ou tem uma nova ideia que não havia considerado antes. É completamente normal; basta atualizar o plano e informar a mudança sendo feita ao stakeholder. Como há restrições de tempo, pode ser necessário remover uma tarefa menos importante do plano existente. Mas, novamente, o plano de análise é útil porque cria uma conversa sobre o que remover, em vez de fazer com que você tente lidar com uma quantidade impossível de trabalho.

#### **Análise de rotatividade de clientes da América do Norte**

August McNamara (amcnamara@company.com), maio de 2020

Objetivo: entender por que os clientes da América do Norte estão se inscrevendo com uma frequência menor do que em outras regiões.

##### **Análise dentro da América do Norte**

1. Há atributos entre os clientes da América do Norte que se relacionam à aquisição dos novos clientes?
  - a. Modelo de regressão dos novos clientes da América do Norte do último mês — gasto e atributos demográficos dos clientes para encontrar a importância
  - b. Estender a seção (a) para comparar entre os clientes adquiridos em cada mês no último ano
2. Como a taxa de aquisição mudou com o tempo?
  - a. Análise da série temporal da taxa de aquisição entre a região
  - b. Série temporal dividida por país/estado e busca de correlações

##### **Comparar a América do Norte a outras regiões**

1. Como os clientes da América do Norte se assemelham com os de outras regiões?
  - a. Modelo linear generalizado com a região como um atributo para modelar a aquisição
  - b. Criar visualização de um mapa global colorido por taxa de aquisição

*Figura 10.3 – Um exemplo de plano de análise.*

## **10.3 Como fazer a análise**

Com a aprovação do plano de análise, é possível começar a análise propriamente dita! O trabalho começa com a importação de dados para que possa manipulá-los e limpá-los. Em seguida, transforma-os repetidamente com a sumarização, agregamento, modificação, visualização e modelagem dos mesmos. Quando os dados estiverem prontos, é hora de divulgar esse trabalho a outras pessoas.

Nas seções seguintes, abordamos brevemente algumas das considerações para ter em mente ao realizar uma análise no ambiente de trabalho. Livros inteiros dedicados a esse assunto também podem ensinar o código para conduzir a análise na linguagem de sua escolha.

### **10.3.1 Importação e limpeza de dados**

Antes de tratar das questões do seu plano de análise, é necessário ter os dados em um local onde possa manipulá-los em um formato que possa utilizar. Geralmente significa ser capaz de carregá-los em R ou Python, mas pode incluir o uso de SQL ou outras linguagens. Quase sempre essa tarefa levará mais tempo do que o esperado. Muitas surpresas podem aparecer durante o processo. Algumas dessas surpresas desagradáveis são:

- Problemas na conexão com os bancos de dados da empresa no ambiente de desenvolvimento integrado (IDE) específico.
- Problemas com tipos de dados incorretos (por exemplo, números como strings).
- Problemas com formatos estranhos de horário (“ano-dia-mês” em vez de “ano-mês-dia”).
- Dados que requerem formatação (talvez cada ID de pedido inicia com “ID-” e é preciso removê-la).
- Registros que estão faltando nos dados.

Pior ainda, esse trabalho parece improdutivo para pessoas não técnicas; não há como mostrar ao stakeholder um gráfico convincente de como conseguiu fazer funcionar um driver de banco de dados, e ele nem entenderia que a manipulação de strings ajuda na solução do problema da empresa. Por isso, por mais entediante que essa tarefa seja, é melhor realizá-la logo para chegar rapidamente na exploração dos dados.

Ao importar e limpar os dados, considere dois objetivos: investir o mínimo de tempo possível naquilo que não for necessário e o máximo que puder no trabalho que irá favorecer os outros. Se tiver uma coluna de dados armazenados como strings e duvidar que precisará dessa coluna, não passe tempo alterando as strings para o formato correto de data e horário. Por outro lado, se achar que precisa dessa coluna, faça o trabalho o quanto antes, pois é necessário ter um conjunto limpo de dados para a análise. É difícil dizer com antecedência o que será útil, mas se perceber que está passando muito tempo em algo, pergunte-se se realmente precisa disso.

Ao importar e editar dados, talvez fique travado vários dias em um único problema, como conectar-se a um banco de dados. Caso se encontre nessa situação, há três opções: (1) peça ajuda, (2) encontre uma maneira de evitar completamente o problema ou (3) continue tentando reparar o problema por conta própria. A opção (1) é ótima se puder fazê-la: uma pessoa mais experiente pode encontrar uma correção rápida, e você pode aprender com isso. A opção (2) também é excelente; fazer algo assim usando um arquivo .csv simples em vez de uma conexão com banco de dados fará a análise desenvolver-se, oferecendo valor à empresa. A opção (3) – continuar tentando de novo e de novo... – deve ser evitada a todo custo. Se passar muitos dias em um único problema, terá a impressão de não ser capaz de oferecer um resultado de valor. Se algo for intransponível, discuta com seu gerente o que fazer, mas não continue tentando nem espere que de alguma forma o problema se resolva por conta própria.

Depois de carregar os dados e formatá-los, é possível começar a usá-los e encontrar dados estranhos. *Dados estranhos* são aqueles fora dos pressupostos básicos. Se estava examinando dados históricos de voo de uma companhia aérea, por exemplo, e encontrou alguns voos que aterrissaram antes de decolar, seria estranho, pois, geralmente, os aviões decolam primeiro! Outra estranheza poderia ser qualquer coisa desde uma loja que vende artigos que têm um preço negativo até dados de fabricação que mostram que uma fábrica fez mil vezes mais artigos do que uma similar. Esses tipos de artefatos estranhos aparecem o tempo todo em dados do mundo real e não há forma de prevêê-los até que sejam examinados.

Se encontrar uma situação com dados estranhos, não a ignore! A pior

coisa que pode fazer é presumir que os dados estão bons e, depois de semanas de trabalho de análise, descobrir que os dados não estavam bons e que seu trabalho foi desperdiçado. Em vez disso, fale com o stakeholder ou com alguém responsável pelos dados que estão sendo utilizados e pergunte-lhe se está ciente do problema. Em muitos casos, eles já sabem e sugerem que você ignore o fato. No exemplo dos dados da companhia aérea, é possível remover apenas os dados para os voos que aterrissaram antes de decolar.

Se a estranheza era desconhecida e poderia comprometer a análise, é necessário investigar formas de recuperá-la. Se você fizer uma análise comparando receita e custos e, estranhamente, metade de seus dados está sem os custos, é preciso ver se pode trabalhar com os custos existentes sozinhos ou com a receita isolada. De certa forma, essa abordagem se torna uma análise dentro de uma análise; você está fazendo uma minianálise para ver se a análise original ainda é viável.

### **10.3.2 Exploração e modelagem de dados**

Durante a exploração de dados e a modelagem da análise, o plano de análise é seguido ponto a ponto para tentar concluir o trabalho. As seções seguintes fornecem um quadro geral que aborda cada ponto.

#### **Use sumarização geral e transformação**

A grande maioria dos trabalhos de análise pode ser concluída com a sumarização e a transformação dos dados. Questões como “Quantos clientes tivemos em cada mês?” podem ser respondidas tomando os dados do cliente, agrupando-os no nível do mês e contando, então, o número distinto de clientes em cada mês. Essa técnica não requer métodos estatísticos ou modelos de machine learning, mas apenas transformações.

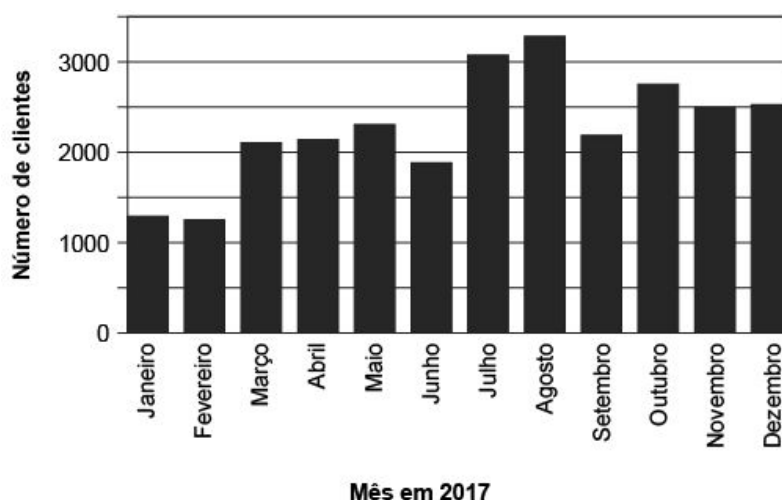
É fácil encarar isso como não sendo realmente uma ciência de dados, porque não requer nada além de muita aritmética, mas, muitas vezes, fazer as transformações da forma correta é bastante valioso. A maioria das outras pessoas na empresa não tem acesso aos dados em primeiro lugar, nem tem a capacidade de realizar as transformações de forma eficaz ou sequer conhece as transformações certas a serem feitas.

Dependendo dos dados, é melhor utilizar alguns métodos estatísticos, como encontrar valores em diferentes níveis de percentis ou calcular um desvio-padrão.

## Visualize os dados ou crie tabelas de sumarização

Depois de fazer as transformações apropriadas, crie visualizações ou tabelas sumarizadas para ver melhor o que está acontecendo com os dados. Continuando o exemplo anterior, se você tivesse o número de clientes a cada mês, daria para gerar um gráfico de barras para ver como mudaram. Esse gráfico pode facilitar a visualização dos padrões nos dados de forma que não se conseguiria pela simples exibição de um quadro de dados em uma tela.

A Figura 10.4 mostra um exemplo de visualização sumarizada mostrando a contagem geral de clientes a cada mês. Com esse gráfico, as pessoas podem ver com facilidade que a contagem de clientes está mostrando uma tendência ligeiramente ascendente.



*Figura 10.4 – Exemplo de tabela sumarizada.*

A visualização a ser escolhida depende significativamente dos dados em mãos. É possível usar um gráfico de linha, um de caixa (box plot) ou qualquer outro dentre muitas outras opções. Também é possível criar tabelas sumarizadas de dados em vez de um gráfico, dependendo do que está tentando entender. Consulte a seção de recursos no fim desta parte do livro para conhecer recursos que podem ajudá-lo a escolher o tipo certo de



gráfico para seus dados. Observe que, ao fazer a visualização, talvez perceba que precisa alterar alguns passos na transformação. Provavelmente voltará e avançará muitas vezes nas etapas.

Como repetirá continuamente os dados por meio de visualizações e transformações, terá de equilibrar a vontade de excluir os dados insignificantes para manter seu código limpo, com a vontade de salvar tudo, caso precise novamente deles. A melhor prática é salvar sempre que possível, desde que (1) seu código antigo não pare de funcionar depois de fazer mais alterações e (2) possa indicar claramente quais resultados são “bons”. Evite manter um código que não funciona na análise ou grandes partes de código sem comentários; essas situações tornam a manutenção do código extremamente difícil. Essa abordagem é melhorada ao utilizar o controle de versões, como o git e o GitHub; ao fazer commits sempre que adicionar novos conteúdos à análise, é possível manter um registro do que foi feito e reverter código que, de repente, pare de funcionar.

## **Crie um modelo conforme necessário**

Se você encontrar padrões em seus dados que sugerem que a modelagem é uma boa ideia, faça isso! Talvez seja bom aplicar um modelo de série temporal aos clientes para prever, por exemplo, os clientes do próximo ano. Ao criar modelos, é melhor produzir resultados e visualizá-los para entender como os modelos são precisos ou úteis. É possível criar gráficos que comparem os resultados previstos com os valores reais ou que mostrem métricas, como medidas de precisão e valores de importância.

Se você criar modelos de machine learning que possam ser usados fora da análise, por exemplo, por ser possível colocá-los em produção (tratado no Capítulo 11), certifique-se de que está isolando o código que constrói o modelo do trabalho geral de análise. Como no futuro você desejará usar apenas o modelo, precisará facilmente ser capaz de retirar esse código do código que faz os gráficos de visualização geral.

## **Repetir**

Complete esses passos para cada ponto do plano de análise. Durante esses passos, talvez você tenha uma nova ideia sobre o que analisar ou perceba

que o que pensava ser uma pergunta sensata não faz sentido. É nesse momento que se deve ajustar o plano de análise e dar continuidade ao seu trabalho.

É provável que os diferentes pontos do plano de análise estejam relacionados; portanto, o código usado em um ponto será repetido em outro. Vale a pena envidar esforços para estruturar o plano de análise para que possa executar o mesmo código repetidas vezes e atualizar instantaneamente uma parte do plano para os outros. O objetivo é construir um conjunto de código que pode ser mantido; é possível facilmente modificá-lo sem gastar horas e horas mantendo o controle do código complexo.

### **10.3.3 Pontos importantes para explorar e modelar**

O trabalho de exploração e modelagem de dados é extremamente dependente do problema que está tentando resolver. As técnicas matemáticas e estatísticas que você usaria para tentar agrupar dados são bastante diferentes daquelas para fazer uma previsão ou tentar otimizar uma decisão. Dito isso, seguir algumas orientações gerais pode fazer a diferença entre uma análise correta e uma análise excelente.

#### **Foque em responder à pergunta**

Conforme discutido na Seção 10.2, é extremamente fácil perder tempo fazendo um trabalho que não condiz com o objetivo. Se estiver analisando pedidos de clientes para ver se pode prever quando um cliente não irá voltar, é possível fazer um modelo de rede neural funcionar corretamente e, então, passar várias semanas ajustando os hiperparâmetros. Se o stakeholder quiser em primeiro lugar apenas uma resposta afirmativa ou negativa sobre a viabilidade do modelo, ajustar os hiperparâmetros para tornar o modelo um pouco mais eficiente não irá ajudá-lo. As semanas que foram passadas no ajuste do hiperparâmetro poderiam ter sido investidas preferivelmente em algo mais relevante.

Ao realizar a análise, é importante manter-se concentrado no plano de análise e responder à questão que a empresa fez. Significa perguntar-se continuamente “É relevante?”. Essa pergunta deve ser considerada cada vez

que fizer um gráfico ou tabela. Se constantemente ficar pensando na importância do que está fazendo, está ótimo. No caso muito mais provável de ocasionalmente pensar que “este gráfico (ou tabela) não é útil”, talvez seja necessário fazer ajustes no seu trabalho. Primeiro, tente parar o que está fazendo e escolha uma abordagem diferente para lidar com o problema. Se estava tentando agrupar os clientes pelos gastos, tente fazer um agrupamento (clustering). Ao adotar uma abordagem dramaticamente diferente, é mais provável que obtenha sucesso do que fazer apenas uma pequena alteração no que está sendo feito. Em segundo lugar, fale com seu gerente ou stakeholder do projeto; pode ser que os dados que estão sendo utilizados não sejam eficazes para resolver o problema em questão.

Nas semanas de análise, você deve construir de maneira estável uma coleção de resultados que são relevantes e (se possível) seguir o plano de análise.

### **Use métodos simples em vez de complexos**

Os métodos complexos são tão animadores! Por que usar uma regressão linear quando se pode utilizar uma floresta aleatória (random forest)? Por que usar uma floresta aleatória quando se pode utilizar uma rede neural? Esses métodos têm demonstrado um desempenho melhor do que uma antiga regressão simples ou agrupamento (clustering)  $k$ -means, além de também serem mais interessantes. Então, quando as pessoas pedem para resolver as questões da empresa com dados, certamente deve usar os melhores métodos possíveis.

Infelizmente, os métodos complexos apresentam muitas desvantagens que não são visíveis ao se focarem apenas na precisão. Durante uma análise, o objetivo não é obter a melhor precisão ou previsão possível, mas, sim, responder a uma questão de modo que alguém da empresa possa entender. Significa que é preciso explicar o motivo pelo qual obteve o resultado. Com uma regressão linear simples, é fácil fornecer gráficos do quanto cada recurso contribuiu para o resultado, ao passo que com outros métodos pode ser muito difícil descrever como o modelo produziu o resultado, o que dificulta que uma pessoa da empresa acredite nesses resultados. Os métodos mais complexos também são mais demorados para serem configurados;

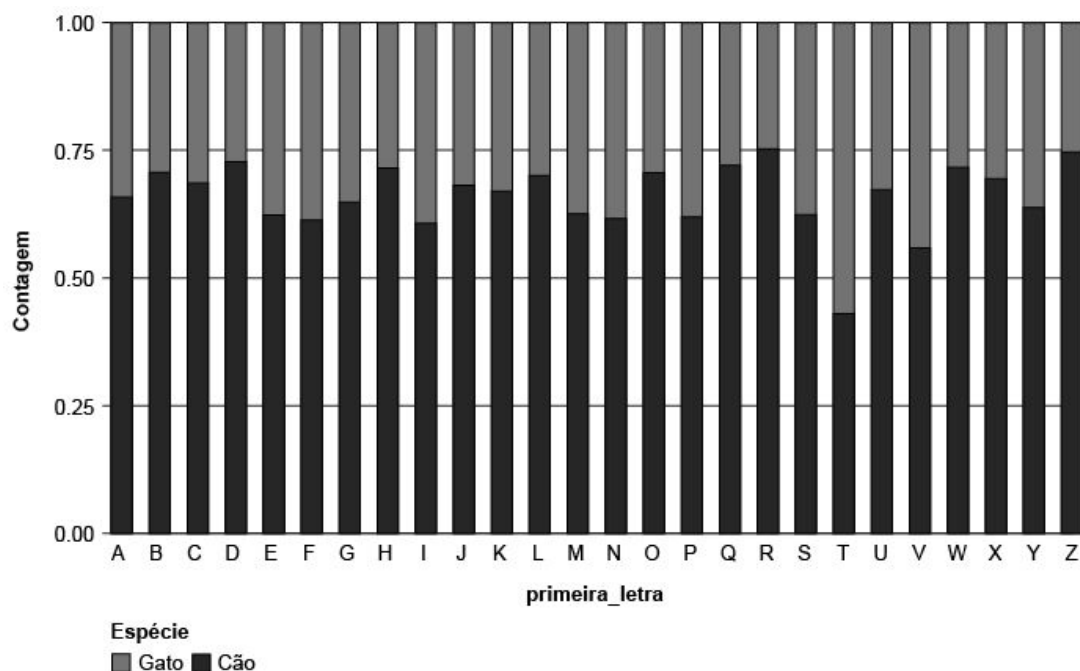
leva algum tempo para sintonizar e executar uma rede neural, enquanto uma regressão linear é bastante rápida.

Assim, quando estiver fazendo a análise, escolha métodos simples o mais frequentemente possível, tanto em modelos como em transformações e agregações. Em vez de podar alguma porcentagem de dados fora da curva, por exemplo, faça uma transformação logarítmica ou tome a mediana em vez da média. Se uma regressão linear funcionar razoavelmente bem, não passe tempo construindo uma rede neural para melhorar um pouco a precisão. Manter métodos simples sempre que possível torna o resultado muito mais fácil para outras pessoas entenderem e para você defender e resolver bugs.

## **Considere os gráficos para exploração x para compartilhamento**

Há duas razões diferentes pelas quais um cientista de dados escolheria visualizar dados: para exploração e para compartilhamento. Quando você está fazendo um gráfico para exploração, o objetivo é ajudar o cientista de dados a entender o que está acontecendo nos dados. Ter um gráfico complexo e mal nomeado é correto, desde que o cientista de dados o compreenda. Quando você está fazendo um gráfico para compartilhamento, o objetivo é que alguém que não conhece muito sobre os dados entenda um ponto específico que um cientista de dados está tentando explicar. Nesse caso, o gráfico deve ser simples e claro para ter eficácia. Ao fazer uma análise, use muitos gráficos exploratórios, os quais não devem ser utilizados para compartilhamento.

Considere um exemplo baseado em dados fictícios sobre nomes de animais de estimação em uma cidade: um cientista de dados quer entender se a letra inicial do nome do animal relaciona-se com a espécie do animal (cão ou gato). O cientista de dados carrega os dados e faz essa visualização, mostrando para cada letra a divisão entre cães e gatos cujos nomes começam com essa letra (Figura 10.5).



*Figura 10.5 – Exemplo de visualização feita durante uma análise antes da limpeza.*

Se examinarmos bem a Figura 10.5, notaremos que a barra T tem um número muito maior de gatos do que de cães – um achado significativo para o cientista de dados. Dito isso, esse gráfico não seria mostrado a um stakeholder; há muita informação nesse gráfico, não ficando claro qual é o resultado à primeira vista.

A Figura 10.6 mostra os mesmos dados no gráfico de uma forma diferente e mais compartilhável. Nessa versão, fica claro que os gatos têm 12% de chance de ter um nome que começa com T, enquanto os cães têm apenas 5% de probabilidade. Agora, esses mesmos dados podem ser compartilhados.



*Figura 10.6 – Os mesmos dados da Figura 10.5, destacando a importância*

*da letra T.*

## **Sempre pronta para compartilhamento**

O resultado da análise pode assumir diferentes formas, sendo que a escolhida é pensada no público-alvo. Se a análise for apresentada a um funcionário da empresa, um slide ou documento editável é frequentemente usado. O PowerPoint ou Word (ou Google Slides ou Google Docs) são uma boa escolha, porque qualquer pessoa pode visualizá-lo (desde que tenha o pacote do Microsoft Office) e pode incluir muitos gráficos, tabelas e descrições em texto. Se a análise for destinada a pessoas com embasamento técnico, é possível apresentá-la em um arquivo HTML do Jupyter Notebook ou no R Markdown. Esses métodos são bons porque normalmente requerem menos trabalho para organizar (ou seja, não é preciso gastar tempo alinhando as figuras em um slide). Se a análise exigir a entrega de muitas tabelas de dados a pessoas do financeiro, o Excel pode ser a melhor escolha. O Excel é uma excelente ferramenta quando o usuário final precisa ter os números nos resultados e fazer cálculos adicionais. Decida no início do processo de análise em que formato espera apresentar os resultados para evitar retrabalho mais tarde.

Dependendo do tamanho do escopo da análise, converse periodicamente com a pessoa para a qual está fazendo a análise e mostre seu trabalho. Essa abordagem evita a maçante situação de passar semanas trabalhando em uma análise isoladamente e, quando for a hora de apresentar os resultados, o stakeholder mostra algo que invalida todo seu trabalho (como: “Você analisou as vendas dos clientes, mas esqueceu-se de considerar as devoluções”). Em uma situação como essa, se a questão tivesse sido apontada no início, teria sido possível evitar descartar muito do trabalho. Além de evitar situações desagradáveis, o stakeholder pode muitas vezes contribuir sugerindo possíveis áreas para se concentrar ou métodos para tentar. De certa forma, ter a confirmação do stakeholder ao longo da análise é similar ao conceito de desenvolvimento de software ágil: continuamente apresentar melhorias no trabalho em vez de liberar uma enorme versão do software.

Tratar disso com frequência com os stakeholders é ótimo, mas os

cientistas de dados muitas vezes negligenciam esse ponto. A desvantagem de ficar confirmando com alguém é que o trabalho tem de estar em um formato para ser mostrado a uma pessoa que não é da área; precisa estar em um nível suficientemente organizado sem que cause constrangimentos. Coisas como gráficos com nomes e significados claros, código com erros mínimos e uma história básica por trás do que está acontecendo são necessárias. Por isso, é fácil para um cientista de dados pensar que: “Vou compartilhar meu trabalho só depois de melhorá-lo, e vou melhorá-lo mais tarde”. Não faça isso! Quase sempre irá gerar mais trabalho a longo prazo. Manter um nível de organização para poder compartilhar seu código gera um produto melhor.

### **Execução com um só botão**

Assim como deve executar apenas um script para carregar e preparar seus dados, sua análise deve ser executada pressionando apenas um botão. Em Python, significa ter um Jupyter Notebook que carrega automaticamente os dados e faz a análise sem erros. Na linguagem R, significa ter um arquivo no R Markdown que carrega os dados, os analisa e envia um arquivo em HTML, documento do Word ou apresentação em PowerPoint.

Ao fazer a análise, evite executar muito código fora do script ou executar seus scripts fora de ordem. Essas práticas tornam mais provável que, ao executar novamente o script inteiro, ocorra um erro. É possível fazer um pouco de programação para essa finalidade, mas certifique-se de que possa executar novamente o arquivo sem erros. Essa prática irá ajudá-lo a manter seus resultados sempre prontos para compartilhar com outras pessoas e garantir que passe menos tempo corrigindo o script no fim da análise.

## **10.4 Encerramento**

Dependendo do stakeholder para essa análise, o resultado do seu código pode ser suficiente para satisfazer a solicitação ou pode ter que ir mais longe e fazer uma versão final. Se uma versão final e organizada for necessária, como uma apresentação em PowerPoint, pode ser necessário reorganizar para seguir as diretrizes de estilo da empresa. Mais importante ainda: precisará elaborar uma narrativa para o documento final, a fim de

que as pessoas que não estavam envolvidas no trabalho possam entender as conclusões do trabalho na sua totalidade, o que foi feito e por quê.

Fazer essa narrativa é o primeiro passo em um bom documento final. Que tipo de história será contada? Como apresentar o problema, explicar como seu trabalho fornece uma solução (ou não) e discutir as próximas etapas? Há muitas formas de criar uma narrativa, mas uma forma simples é pensar em como explicar o trabalho em voz alta a uma pessoa que não o conhecia. Pense sobre a história que contaria e tente contá-la no documento. Faça repetidamente as seguintes perguntas: “O que estou mostrando estará compreensível pelo meu público?” e “O que posso fazer para melhorar?”. Por fim, chegará a um ponto em que estará muito satisfeito com o conteúdo.

De igual modo, precisará adicionar texto ao documento – normalmente para explicar a narrativa que tem ou por que cada gráfico vale a pena ser compartilhado. Mais uma vez, tente torná-lo compreensível para alguém que não tem o contexto do trabalho. O texto responde à pergunta: “De que forma o que estou mostrando é útil para a empresa?”. Empresas diferentes têm padrões distintos sobre quanto de texto incluir; algumas querem descrições detalhadas explicando tudo; outras, ficam satisfeitas com algumas palavras. Tente pecar por explicar demais, pois será possível para cortar conteúdo depois.

Quando pensar que seu material está pronto, é bom um colega revisá-lo buscando pequenos erros antes de apresentá-lo ao stakeholder. Considere ter alguém da equipe que esteja familiarizado com o contexto do trabalho para verificar se tudo faz sentido. Dependendo da empresa, o gerente pode exigir que essa etapa seja realizada com ele para aprová-la.

### **10.4.1 Apresentação final**

Quando obtiver a aprovação da análise pelo seu gerente, agende uma reunião com o stakeholder para apresentar a análise pessoalmente. Nessa reunião, explique cada componente, descrevendo o que foi feito, o que aprendeu e o que optou por não examinar. Tendo passado tanto tempo com os dados gerando a análise, você deve estar bastante confortável para explicar e responder perguntas.

Dependendo do stakeholder, é possível receber muitas perguntas durante a



apresentação ou a pessoa aguardará até o fim para questionar. As perguntas podem variar entre tranquilas e curiosas (“Por que usou o conjunto de dados X em vez de Y?”) e críticas e preocupadas (“Por que esses resultados não se alinham com o trabalho da outra equipe? Existem erros no seu código?”). Lidar com perguntas é, de muitas maneiras, a mesma coisa que respondê-las em entrevistas de emprego (Capítulo 7): seja honesto com o que sabe e com o que não sabe. Não há problema em dizer que precisa examinar algo. Tanto quanto possível, explique seu raciocínio (“Usamos o conjunto de dados X porque cobria o período que nos importava”) e quando não souber algo (“Não tenho certeza por que não se alinham com a outra equipe; vou averiguar isso”). Dito isso, na maior parte das vezes essas reuniões são calmas e sem conflitos!

Não importa quão boa seja sua análise, inevitavelmente perguntarão algo como: “Bem, e sobre \_\_\_\_\_?”, em que o espaço em branco é algo não analisado. Alguém pode perguntar: “Bem, e se usar apenas os dados do último mês na análise?”. É natural fazerem isso por causa da natureza da ciência de dados: sempre há mais maneiras de dividir os dados e ideias sobre o que pode ser útil. É comum em situações nas quais a análise se revelou inconclusiva. Nessas situações, a pessoa que fez o pedido quer ter a esperança de que algo pode de repente se provar conclusivo.

Como cientista de dados, a melhor coisa a fazer nessas situações é tentar rejeitar delicadamente essas solicitações. Embora ocasionalmente sejam úteis, elas podem simplesmente acabar levando à nenhuma conclusão nova, tendo apenas feito você perder dias tentando trabalhar nelas. Como cientista de dados, você tem que saber o que teria chance de ser valioso, mas, se achar que algo não seria útil, dá para explicar essa conclusão. Muitas vezes, quando está fazendo uma análise, a questão da empresa que está tentando resolver é tão abstrata que nunca poderia dar uma resposta verdadeiramente definitiva. E assim como quando estava fazendo a análise e tentava evitar testar um método depois do outro para encontrar um resultado, após a análise, é preciso saber em quando parar.

### **10.4.2 Como salvar seu trabalho para a posterioridade**

Quando a análise final for apresentada e aprovada, solicitarão que passe

rapidamente para o próximo trabalho, por exemplo outra análise. Antes disso, entretanto, seguir alguns pequenos passos facilitará muito sua vida no futuro. Há uma boa chance de que em algum momento, meses ou anos depois, solicitem que uma análise seja refeita com dados mais recentes. Se você passar algum tempo documentando seu trabalho, repetir essa análise será muito mais fácil. Os passos são:

- *Confirme se pode executar novamente toda a análise.* Antes, discutimos como fazer da sua análise uma execução com um só botão; a essa altura, faça uma última verificação para confirmar se a análise ainda funciona.
- *Comente seu código.* Como pode não ter revisto seu código há anos, até mesmo comentários curtos podem ajudá-lo a lembrar-se de como usar ou modificar seu código.
- *Adicione um arquivo README.* Um arquivo README é um documento de texto simples que abrange para o que a análise foi feita, por que foi feita e como executá-la.
- *Guarde seu código com segurança.* Se estiver usando o git e o GitHub, já terá feito isso, mas, se não tiver, considere como alguém poderia acessar o código daqui a muito tempo.
- *Certifique-se de que os dados estão armazenados com segurança.* Verifique se todos os arquivos de dados estão salvos em um local seguro que não seja seu computador, como nos serviços na nuvem (OneDrive, uma unidade de rede compartilhada ou AWS S3, por exemplo). Além disso, os conjuntos de dados armazenados em bancos de dados devem ser preferencialmente verificados para garantir que não sejam excluídos.
- *O resultado é armazenado em um local compartilhado.* A forma mais comum de as pessoas compartilharem análises é como anexos de email, mas não é uma boa forma de arquivá-las. Coloque os resultados em um lugar que outras pessoas da equipe e de outros setores da empresa possam acessar.

Quando finalizar essas etapas, finalmente você poderá dizer que a análise foi concluída de verdade. À medida que faz cada vez mais análises, encontrará os métodos e técnicas que funcionam melhor, e ficará melhor e mais rápido neles.

## 10.5 Entrevista com Hilary Parker, cientista de dados na Stitch Fix

Hilary Parker é doutora em bioestatística pela Faculdade de Saúde Pública da Johns Hopkins Bloomberg, e trabalha na Stitch Fix, um serviço de estilização pessoal online, onde cria modelos de machine learning para ajudar a sugerir roupas para os clientes. Antes, ela era analista de dados sênior na Etsy.

### Como pensar em outras pessoas ajuda na análise?

Praticamente todas as análises que começo tento entender “Quem quer o quê?”. Por exemplo, o trabalho foi solicitado porque o gerente de produto precisa tomar uma decisão e não sente que conseguirá até ter essa análise de um experimento? Será que estamos tentando conduzir uma visão estratégica e, para que as pessoas se sintam confortáveis com ela, precisamos mostrar que acreditamos que isso faria X dólares ao longo de Y anos? Eu me sento e falo com os eventuais usuários da análise para compreender o contexto.

Na apresentação, o mais importante é entender o público, onde essas pessoas estão e quais são seus objetivos. Querem ou não entender cada detalhe? O que seria mais atraente para elas? Se parecem ansiosas por mais informações, é possível oferecer mais detalhes estatísticos, mas, se não estiverem engajadas, é melhor reduzir os detalhes.

### Como estruturar as análises?

Penso que é importante estruturar a análise de uma forma acessível. Faço um breve resumo no topo, mas não faço gráficos complicados, porque a maioria das pessoas não consegue absorvê-los rapidamente. Também não faço anotações no estilo “fluxo de consciência” para uma análise (o que vejo muitas pessoas fazendo no ambiente empresarial). Os comentários nas anotações são semelhantes a um texto; então, é possível adicionar mais e mais comentários. No fim, acabam apresentando algo como: “Aqui foi onde comecei e aqui foi onde terminei”. Mas é melhor trocar a ordem: “Aqui está a Conclusão e, no Apêndice, dá para ver onde comecei”. Tenha em mente a ideia de que alguém irá lê-lo e o que será mais fácil de produzir

rapidamente pode não ser o mais legível. Foco tanto no formato final que ele faz parte do processo. Não tenho de traduzir uma grande anotação e deixá-la apresentável; deixo-a sempre apresentável.

## **Como deixar a versão final mais apresentável?**

Penso que o tema das cores é uma maneira deixar a versão final mais apresentável. Muitas empresas têm um tema já definido; a Stitch Fix tem um tema de cores na marca. Temos modelos ggplot2 que importam as cores da nossa paleta, o que é realmente eficaz, pois faz com que o pessoal da empresa sintam-se familiarizado. Fazemos o mesmo com as apresentações do Google Slides. Há modelos do Google Slides que as pessoas usam porque parece bons.

Também penso: “Não exagere”. Um dos meus primeiros projetos na Stitch Fix foi lançar nossa linha de negócios de tamanhos grandes (plus-size). Fizemos uma análise rápida necessária para entender se estávamos enviando os tamanhos certos. Passei tanto tempo construindo meu pequeno sistema quanto a forma de apresentar a análise. Fiquei animada em desenvolver um site reproduzível que atualizaria de maneira dinâmica a cada  $x$  horas para mostrar o que estava sendo alterado. Mas, no fim, as pessoas com quem trabalhava não deram muita atenção a isso. Fiquei muito animada em construir o site em vez de confirmar com os colegas. É fácil animar-se demais com a estética da análise. Faça o que for necessário, mas não demais.

## **Como lidar com as pessoas que pedem ajustes a uma análise?**

Recentemente tenho lido muito sobre design thinking, o que acontece o tempo todo no contexto de design. A atitude que tenho tomado para mim mesma é a de que as pessoas são ruins na comunicação e não vão pensar de forma abstrata. No mundo do design, a pessoa dirá o que quer, e você não pode levar isso literalmente. Precisa ajudá-los a enquadrar o problema. Isso faz parte do valor agregado de um designer: pensar no problema de forma holística, enquadrar de forma sistemática e de diferentes maneiras até fazer sentido.

Penso que os cientistas de dados e os estatísticos são assim também.

Alguém pedirá algum recurso porque está tentando tratar de um problema, e essa é uma maneira de expressá-lo. Mas é preciso entender que problema é esse. Está dizendo que não quer tomar essa decisão? Está causando hesitação? Qual será o resultado final dessa situação? Como cientista de dados, quase sempre está interagindo com um dos consumidores. Você não tem que fazer sempre apenas o que dizem, mas descobrir o que a pessoa está tentando realmente dizer. Qual é a causa principal do que estão dizendo? É algo que uma análise consegue abordar? Muita coisa pode estar acontecendo, e é importante ter uma perspectiva geral sobre toda a situação, em vez de apenas repetir até o fim.

## **Resumo**

- As análises são documentos que destacam conclusões e englobam recursos importantes de uma aplicação da ciência de dados para resolver um problema da empresa. Elas são cruciais para os cientistas de dados.
- Uma boa análise requer a compreensão do problema da empresa e como os dados podem resolvê-lo.
- Ao fazer a análise, sempre pense no objetivo final, use métodos simples com visualizações claras e esteja pronto para compartilhar o trabalho.
- A gestão do processo de criação da análise é importante para manter o trabalho focado no objetivo e garantir que tenha uma finalidade clara.

# CAPÍTULO 11

## Como implantar um modelo na produção

Este capítulo abrange:

- Construir um modelo de machine learning para uso na produção
- Compreender o que são APIs e sua utilidade
- Como implantar um modelo de machine learning

Este capítulo trata dos conceitos essenciais do trabalho de um engenheiro de machine learning – alguém que cria modelos de machine learning e os implementa para uso da empresa. Se seu trabalho, em vez disso, envolve a criação de análises e relatórios, é fácil preocupar-se com esse material. Não precisa! A lacuna entre o cientista de decisão e o engenheiro de machine learning é menor do que parece, e este capítulo será uma introdução útil aos conceitos.

Às vezes, o objetivo de um projeto de ciência de dados não é responder a uma questão com dados, mas criar uma ferramenta que use um modelo de machine learning para fazer algo útil. Embora possa realizar uma análise para entender quais itens as pessoas tendem a comprar juntos, é uma tarefa diferente fazer um programa que recomende o melhor item no site para um cliente. O trabalho de pegar um modelo de machine learning e de fazê-lo de modo que possa ser usado por outros setores da empresa, como no site ou no call center, tende a ser complexo, além de envolver cientistas de dados, engenheiros de software e gerentes de produto.

Neste capítulo discutiremos o modo de pensar sobre como construir modelos que fazem parte de um produto e como tirá-los do seu computador e colocá-los em um lugar onde possam funcionar.

Duas pequenas observações antes de mergulharmos nesse tópico:

- Como a tarefa de criar código executada na produção é bastante técnica, este capítulo é mais técnico do que os demais. Como queremos que esses tópicos sejam de fácil compreensão para quem tem menor familiaridade com os conceitos de desenvolvimento de software, vamos nos concentrar mais nos conceitos e nas ideias do que nas especificações técnicas.
- Uma vez que estamos nos concentrando mais nos conceitos, às vezes serão feitas exposições gerais que podem não ser 100% verdadeiras. Essa decisão é intencional, feita para ajudar na leitura. Se estiver familiarizado com esses tópicos e consegue pensar em um argumento contrário a algo que escrevemos – você provavelmente está correto!

## **11.1 O que significa implantar na produção?**

Quando as pessoas dizem “implantar na produção” significa o mesmo que colocar o código em algum tipo de sistema que permita sua execução de forma contínua, normalmente como parte de um produto voltado ao cliente. Implantar é um verbo que significa mover o código para um sistema diferente, enquanto produção é um substantivo, ou seja, o lugar onde o código que faz parte de um produto é executado. O código que está em produção precisa ter capacidade de funcionar com erros ou problemas mínimos, pois, se o código parar de funcionar, os clientes notam.

Embora os desenvolvedores de software tenham colocado o código em produção há décadas, ele está se tornando cada vez mais comum para os cientistas de dados, especificamente para os engenheiros de machine learning, para treinar um modelo de machine learning e colocá-lo em produção também. Treinar um modelo de machine learning para ser colocado em produção assemelha-se a treinar um modelo como parte de uma análise, mas há mais passos após treinar o modelo para prepará-lo para a produção. Com frequência, o ato de construir o modelo para a produção começa com uma análise. Primeiro, é preciso entender os dados e obter o aval da empresa; então, você pode pensar sobre a sua implantação na produção. Assim, os dois atos estão bastante interligados.

Para ter uma noção melhor do que significa a implantação na produção, a

seguir apresenta-se um pequeno exemplo. Suponha que um stakeholder de uma empresa pensa que muitos clientes estão indo embora e pede a um cientista de dados para fazer uma análise da rotatividade da clientela. Como parte da análise, o cientista de dados constrói um modelo e mostra que existem vários indicadores-chave de rotatividade. O stakeholder aprecia a análise e dá-se conta de que se os agentes de atendimento ao cliente que trabalham na central de atendimento soubessem quais clientes provavelmente entrarão em rotatividade, poderiam lhes oferecer descontos para tentar evitar que saíssem.

A essa altura, o cientista de dados precisa colocar o modelo em produção. O modelo que estava no computador do cientista de dados deve ser executado toda a vez que um cliente ligar para o suporte, a fim de avaliar a chance de rotatividade. No computador, o modelo levou alguns minutos para avaliar muitos clientes ao mesmo tempo, mas, na produção, esse modelo tem de ser executado em um único cliente quando esse ligar, obtendo os dados do cliente de outros setores da empresa e usando esses dados para uma classificação (rating).

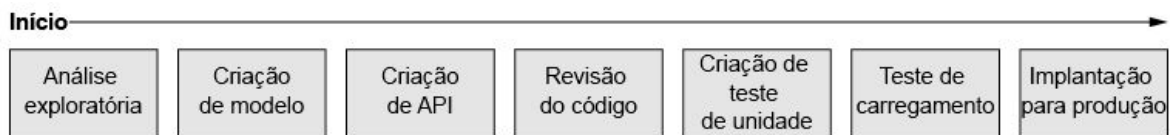
A maioria dos modelos de machine learning de produção é similar: precisa trabalhar em tempo quase real para fazer uma previsão ou classificar algo com base nos dados fornecidos. Exemplos famosos são o modelo de recomendação de filmes da Netflix, o qual prevê os filmes que uma pessoa gostaria; o modelo de reconhecimento facial do Facebook, que tira uma imagem, encontra rostos e corresponde esses rostos a identidades; e o modelo de preenchimento automático do Gmail da Google, que pega o texto à medida que é escrito e prevê a próxima palavra.

Os modelos usados para a produção precisam seguir vários passos substanciais. Primeiro, os modelos precisam ser codificados para lidar com qualquer cenário que possa acontecer quando o código estiver ativo, de modo que sejam menos propensos a erros. Quando você está fazendo uma análise, pequenas quantidades de dados estranhos podem ser filtradas e não alimentadas no modelo sem interromper os resultados da análise. Em um modelo de produção, o código precisa ser executado independentemente de quão estranhos os dados inseridos sejam. Provavelmente, é bom para fins de análise, se um modelo de processamento de linguagem natural falhar



quando executado em um emoji, por exemplo, pois você pode simplesmente ignorar os dados inseridos com emojis. Para um modelo de produção, se o código se quebrar quando um emoji aparecer, pode ser que o produto que o modelo de machine learning suporta também se quebre. Imagine o que aconteceria se a página web do Gmail travasse cada vez que um emoji fosse digitado. Os modelos de produção precisam ser feitos para lidar com casos estranhos ou o código precisa corrigir casos especiais e estranhos antes que cheguem ao modelo.

Os modelos de produção também devem ser passíveis de manutenção. Como estão sendo continuamente utilizados em produtos, algumas vezes têm de ser retreinados em dados mais recentes – ou codificados de forma que se retrainem automaticamente. Precisam de maneiras para monitorar o quão bem estão funcionando para que as pessoas na empresa possam dizer se já não estão funcionando bem ou se, de repente, param de funcionar completamente. E, como é possível que estejam funcionando durante anos, precisam ser codificados de forma que sigam as normas de outros modelos e possam ser atualizados ao longo do tempo. Codificar um modelo em uma linguagem de programação arcaica, que poucas pessoas saibam é ruim para uma análise e catastrófica para um modelo de produção. Consulte a Figura 11.1 para ver um exemplo de processo de criação e implantação de um modelo de machine learning na produção.



*Figura 11.1 – Exemplo do processo de criação de um produto de machine learning na produção.*

O restante deste capítulo abrange três conceitos: como criar um modelo de machine learning adequado para produção, como implantá-lo na produção e como mantê-lo em funcionamento ao longo do tempo.

## **Produção para diferentes tipos de cientistas de dados**

O quanto é preciso pensar sobre sistemas de produção varia muito com o tipo de cientista de dados que você é:

- *Engenheiro de machine learning* – praticamente é tudo o que você faz. Quando se sentir confortável na função de engenheiro de machine learning, também estará à vontade com tudo o que é discutido neste capítulo.
- *Analista* – um analista pode ter que lidar com sistemas de produção, dependendo da complexidade dos relatórios. Se a equipe de análise estiver continuamente gerando relatórios consistentes, é melhor produzir os sistemas de relatórios. Fazendo com que os relatórios se atualizem automaticamente, a equipe de análise é livre para fazer outro trabalho. É especialmente verdadeiro para dashboards, em que a expectativa é que os sistemas estejam em produção, atualizando-se por conta própria.
- *Cientista de decisão* – como o trabalho de um cientista de decisão é principalmente ad hoc, não há tantas oportunidades para criar sistemas de produção. Mas se os modelos criados pelos cientistas de decisão forem entregues a engenheiros de machine learning para produzir, será muito importante ter um melhor entendimento dos sistemas de produção. Os cientistas da decisão também podem criar ferramentas interativas para a empresa com base em bibliotecas, como Shiny ou Dash, que precisam ser implantadas e mantidas em sistemas de produção.

## 11.2 Como criar o sistema de produção

Um sistema de produção baseado em um modelo de machine learning inicia com as mesmas etapas de fazer um modelo de machine learning para uma análise; é preciso encontrar os dados apropriados, definir as características, treinar o modelo e receber o aval da empresa. Quando essas etapas forem concluídas, mais trabalho terá de ser feito:

1. O modelo precisa ser convertido para um formato que outros programas possam usar. Normalmente, é feito com a criação de código que permite que o modelo seja acessado como uma API de outros sistemas da empresa como se fosse um site.

2. O modelo deve ter código para lidar com muitas entradas possíveis. Assim, garante-se que não ocorrerão problemas com o modelo, caso sofra uma entrada inesperada, e também terá o mínimo de tempo de inatividade possível. Essa etapa requer a adição de testes ao modelo a fim de garantir que esteja lidando corretamente com todos os dados que poderiam ser processados.
3. O modelo é implantado em um ambiente de teste para garantir que funcione corretamente. A API é testada para garantir seu funcionamento e que possa lidar com a quantidade de tráfego que a atingirá quando estiver ativa.

Quando todas essas etapas forem concluídas, o modelo será finalmente implantado em um ambiente de produção.

### **11.2.1 Coleta de dados**

Ao coletar dados para treinar um modelo para uma análise, é preciso encontrar um conjunto adequado de dados históricos que contenha um bom indicador. Embora também seja necessário para um modelo de produção, muitas vezes não é suficiente, pois o componente em tempo real do modelo precisa ser levado em conta. Considere o exemplo anterior no qual uma empresa precisa de um modelo de produção para prever em tempo real se um cliente irá deixá-la. Se um modelo de rotatividade de cliente for usado para uma análise, uma coleta histórica de atributos do cliente (número de compras, anos desde a primeira compra e assim por diante) que foi feita há alguns meses seria ótima para um modelo. Como o modelo de produção precisa prever em tempo real se um cliente irá embora, o código precisará ser programado para descobrir de alguma maneira quais são os atributos do cliente quando o modelo for chamado. Se o cliente #25194 estiver ligando para o suporte ao cliente, o código para o modelo precisa saber naquele momento exato quantos pedidos o mesmo fez para que o modelo possa determiná-lo.

A diferença entre utilizar dados históricos para treinar um modelo e alimentar os dados em tempo real do modelo durante a execução pode ser drástica. Por motivos técnicos relacionados com a forma como os dados são coletados, pode haver um atraso de horas ou dias antes de os dados serem

colocados em um banco de dados ou local de armazenamento a que um cientista de dados possa acessar. Como alternativa, pode haver situações nas quais os dados estejam disponíveis em tempo real, mas os valores não são armazenados historicamente. Pode haver uma maneira de perguntar se um cliente está localizado internacionalmente no momento, por exemplo, mas os dados sobre se o cliente algum dia foi internacional não foram armazenados.

Quando estiver procurando dados para criar um modelo para produção, considere quais dados serão necessários em tempo real quando o modelo for executado. Os dados serão novos para utilização? Podem ser acessados por meio de uma conexão de banco de dados ou de algum outro método que alguém pode configurar? Não é incomum que os projetos de machine learning falhem devido a problemas no conjunto de dados.

### **11.2.2 Como construir o modelo**

Quando tiver um conjunto de dados adequado, você pode começar a construir um modelo de machine learning. Esse tópico é amplo; se quiser aprender a como construir um modelo de machine learning, desde a engenharia dos recursos ao treinamento do modelo até a validação, há vasta bibliografia e recursos disponíveis na internet. Dito isso, ao criar um modelo específico para a produção, é bom ter algumas coisas em mente.

#### **Prestar especial atenção ao desempenho do modelo**

Como outros sistemas irão depender do seu modelo para trabalhar de forma eficaz em quaisquer dados que esses sistemas optem por passar para ele, é preciso compreender como o modelo funcionará em todos os casos. Suponha que esteja fazendo um modelo de machine learning como parte de uma análise para entender em quais produtos os clientes estão interessados, como um modelo que prevê o produto que um cliente irá comprar a seguir. Se o modelo que construiu corretamente previu 99% das compras, mas 1% previu que o cliente encomendaria tatuagens temporárias do Nicolas Cage, o modelo seria um enorme sucesso. Ao compreender a grande maioria dos clientes, é possível ajudar a empresa a tomar uma decisão consciente de marketing. Se, por outro lado, esse modelo fosse implantado para mostrar

os produtos recomendados no site da empresa, esse 1% poderia ser catastrófico e fazer perder os clientes (ou, pelo menos, os clientes que não apreciam o carisma incomparável de Nicolas Cage!). O que acontece nas margens é realmente importante nos sistemas de produção, mas não nas análises.

## **Como construir um modelo simples**

Quando o modelo for implantado na produção e estiver em execução para que os clientes interajam com ele, inevitavelmente se deparará com uma situação na qual o modelo faz algo estranho e é preciso entender por quê. Se for um modelo de rotatividade para o cliente, ele poderá prever que cada cliente no Alaska entrará em rotatividade por algum motivo desconhecido. Ou um motor de recomendação em um site de produtos para usar ao ar livre pode recomendar apenas caiaques. Nesse momento, terá de descobrir o que está acontecendo e se o modelo precisa ser alterado de alguma forma.

Se usar um modelo simples, como uma regressão linear, ele deve ser bastante simples para rastrear o cálculo feito para a predição. Se utilizar um método complexo, como um conjunto ou um modelo reforçado, torna-se muito mais difícil compreender o que está acontecendo. Embora deva usar um modelo que seja complexo o suficiente para resolver seu problema, use tanto quanto possível o modelo mais simples e aceitável, mesmo perdendo em precisão.

Uma história interessante do mundo real envolve o Prêmio Netflix. A empresa realizou um concurso para ver se uma equipe conseguiria criar um algoritmo que melhorasse os resultados da recomendação de filmes em 10% e, em 2009, concedeu um prêmio de 1 milhão de dólares. Como indicado em um artigo da *Wired* (<https://www.wired.com/2012/04/netflix-prize-costs>), a Netflix acabou nunca usando o algoritmo vencedor. O algoritmo foi um método de conjunto que combinava muitos modelos, mas a complexidade de engenharia da execução e resolução de bugs era tão alta que sua utilização não valeria a pena o aumento na precisão. Apesar do fato de a Netflix ter pagado um preço elevado para ter um modelo preciso, a empresa percebeu que existem coisas mais importantes do que a precisão. Se a Netflix, com seu exército de cientistas e engenheiros de dados, não mantém

um modelo altamente complexo, seria difícil dizer que muitas empresas o necessitam.

### 11.2.3 Como atender modelos com APIs

Desde que este livro foi escrito, a maioria dos modelos de machine learning é atendida como interfaces de programação de aplicativos (APIs). Significa que o código do modelo de machine learning pode ser executado em um sistema de computador e que outros sistemas podem se conectar a ele quando precisarem que o modelo seja executado em seus dados. Se uma empresa tiver um sistema que execute o site de compras e pretender adicionar um modelo de machine learning para oferecer um desconto caso preveja que o cliente se descadastrará, em vez de tentar obter o código de machine learning dentro do código do site, seria possível configurar um segundo sistema que tenha o modelo de machine learning, podendo ser periodicamente consultado pelo site.

O conceito de dividir partes diferentes de um sistema em microsserviços é uma fonte de controvérsia na engenharia de software e tratado com profundidade em muitos livros. Para os cientistas de dados, o conceito importante é a configuração de um sistema de computador que executa apenas o modelo para que outros sistemas possam usá-lo.

As APIs modernas são feitas com serviços da web (web services), chamadas coloquialmente de REST APIs. Uma REST API é basicamente um site pequeno, mas, em vez do site retornar HTML para ser renderizado em um navegador, ele retorna dados, em geral como texto formatado. Essas requisições usam o protocolo HTTP, que é o mesmo que os navegadores da web usam (e é por isso que os endereços do site começam com `http://` ou `https://`). Uma API meteorológica, por exemplo, poderia ser configurada de modo que, ao acessar a URL [http://exampleweather.com/seattle\\_temperature](http://exampleweather.com/seattle_temperature), o site retornaria a temperatura em graus de Seattle (45). Para uma API de modelo de machine learning, seria desejável ir para um site em especial e obter uma previsão. No caso de um modelo de machine learning, em algum lugar na rede da empresa, poderia haver um local que prediz se um cliente irá embora. Um site como `http://internalcompany.com/predict?customer=1234` retornaria

um número entre 0 e 1, representando a probabilidade de o cliente ir embora.

Projetar a API envolve tomar decisões como que URLs retornam quais dados e quais os tipos de requisições usar. Tornar o projeto compreensível aos usuários é uma parte importante de fazer as pessoas usarem-no realmente, ou seja, deve-se pensar na interface assim como em criar o modelo.

Ter um modelo de machine learning executado como um serviço web da API é ótimo por vários motivos:

- Como se trata de uma API da web, qualquer coisa ou pessoa na empresa pode usar o modelo, incluindo sistemas em tempo real e outros cientistas de dados que executam análises. O mesmo modelo de previsão de rotatividade que está sendo usado no site pode ser consultado para uma análise por uma pessoa que está fazendo ciência da decisão.
- Uma vez que funciona como um site, quase toda a tecnologia moderna pode ser conectada ao modelo, independentemente da plataforma técnica. Se o modelo estiver escrito em R, ele ainda poderá ser usado por um site escrito em Node.js ou por outro analista que usa Python. Além disso, como é hospedado como o próprio site dele, há menos chances de que outros produtos sejam removidos por causa dele, caso o modelo deixe de funcionar por algum motivo. Se o site de compras da empresa usar o modelo e de repente não conseguir se conectar, o site de compras ainda pode seguir funcionando.

### **11.2.4 Como construir uma API**

As APIs são ótimas para modelos de machine learning, mas é necessário algum código adicional para elas. Tanto R quanto Python têm pacotes – Plumber e Flask, respectivamente – que fazem essa codificação para você. Quando executa um script R ou Python com esses pacotes, esse script assume sua função e o expõe a um endpoint no computador. Seria possível especificar que ir para a URL *http://yourwebsite.com/predict* executará sua função de machine learning em R ou Python e, em seguida, retornar qualquer que seja o resultado da função. E, então, poderia entrar em um navegador da web para chamar seu código! Supondo que você esteja

executando esse código em um computador portátil ou desktop, se configurar sua máquina para permitir tráfego externo ajustando o firewall, outras pessoas poderão chegar na sua API. Entretanto, quando parar de executar o programa de hospedagem API (R ou Python), ninguém será capaz de executar seu modelo.

Embora tanto R quanto Python facilitem atender a um modelo como uma API, decisões de design precisam ser tomadas, como quais dados precisarão ser passados como entrada para o modelo em uma requisição (request) de API. Suponha que você esteja fazendo um modelo para prever a probabilidade de evasão de um cliente, a quantidade que gastou com a empresa e o número de vezes que ligou para o atendimento ao cliente. Um possível design de API seria para uma pessoa fazer uma requisição com o ID exclusivo do cliente na URL. Para pesquisar o cliente com o ID 1234, por exemplo, dá para acessar *[http://yourwebsite.com/predict?customer\\_id/1234](http://yourwebsite.com/predict?customer_id/1234)*. Outra opção seria os usuários terem que pesquisar todas as informações do cliente e, em seguida, incluir essas informações como corpo de uma requisição. Assim, para um cliente com uma duração de 1,7 anos, um gasto total de US\$ 1.257,00 e três ligações para o centro de atendimento, é possível enviar uma requisição para *<http://yourwebsite.com/predict>* onde a requisição do corpo fosse {"estabilidade": 1,7, "gastar": 1257, "ligações": 3 }.

Ambas as opções são válidas para um design de API, mas uma requer que a API faça todo o trabalho de pesquisar os detalhes do cliente, enquanto a outra faz com que o usuário da API pesquise os detalhes. Geralmente não é uma boa ideia tomar essas decisões sozinho; quanto mais você puder envolver as pessoas que poderão usar sua API, maiores são as chances delas ficarem satisfeitas.

Depois de ter projetado sua API, fale com os usuários e mostre como ela funciona. Eles podem dar feedback sobre o design. Se possível, também compartilhe a documentação sobre a API com elas.

## **Sobre o Plumber: um pacote R para atender o código R como uma API (Jeff Allen)**



*Jeff Allen trabalha na RStudio e é o criador do pacote R Plumber, o qual permite que as pessoas criem APIs em R para que os modelos de machine learning possam ser usados em toda uma empresa.*

O Plumber não começou com uma empresa que decidiu financiar e colocar muitos engenheiros para criá-lo; ele teve uma origem mais humilde. Em 2012, estava trabalhando com um grupo de bioestatística em um centro de pesquisa. Esse grupo usava R para grande parte da análise e me levou para ajudar na criação de um software melhor. No início, tínhamos três públicos diferentes:

- Outros bioestatísticos que usavam R e queriam aproveitar ou avaliar os métodos que tínhamos desenvolvido.
- Usuários não técnicos, como médicos, que apenas precisavam dos resultados de nossa análise para responder a questões como: “Qual medicamento seria mais eficaz para esse paciente?”
- Usuários técnicos que queriam aproveitar nossa análise, mas não estavam interessados ou não tinham os recursos computacionais para executar nosso código R.

O primeiro público foi o mais fácil; R vem com um sistema de empacotamento robusto que permite agrupar seu código e seus dados em pacotes que podem ser compartilhados e usados por outros. O segundo público era um pouco mais difícil de atender na época. Hoje, o Shiny é a solução óbvia para esse problema e oferece uma maneira conveniente para que os usuários de R continuem trabalhando em R para construir aplicativos web ricos e interativos passíveis de serem consumidos por um público não técnico.

Esse terceiro público continuou difícil de se lidar. Alguns usuários já tinham um aplicativo existente escrito em outra linguagem, como Java, mas queriam invocar algumas de nossas funções R a partir do serviço deles. Outros tinham um pipeline simples e automatizado e queriam aproveitar alguma função de R computacionalmente intensiva que tínhamos definido. Em todos esses casos, o que na realidade queriam era algo que poderiam chamar remotamente e confiar em nós para fazer internamente todo o processamento R e, em seguida, enviar-lhes o resultado. Em suma, queriam uma API remota para R.

Somente anos mais tarde tive a real oportunidade de começar a trabalhar no pacote que se tornaria o Plumber, mas levei essa motivação comigo. Há um grupo de pessoas na maioria das empresas que não conhece R, mas que se beneficiaria da análise de que seus cientistas de dados estão criando em R. Para muitos, uma interface programática e estruturada é necessária, e as APIs oferecem uma solução elegante. Felizmente, os autores do pacote Shiny já haviam resolvido todos os problemas difíceis em torno da criação de um servidor web com bom desempenho que poderia ser usado por pacotes R para atender às requisições HTTP. O que faltava era criar uma interface pela qual os

usuários pudessem definir a estrutura e o comportamento da API.

Minha esperança era que o Plumber oferecesse uma solução a esse público técnico de modo que os usuários programáticos pudessem se beneficiar de R de maneira tão eficaz quanto os outros públicos aqui mencionados. Como acompanhei o crescimento do Plumber nesses anos, tanto em termos de funcionalidades quanto de utilização, penso que ele foi aprovado pelos usuários de R. Como o R pode ser convenientemente superado por uma API, agora ganhou um lugar ao lado de outras linguagens de programação que são mais familiares às empresas de TI tradicionais. É divertido ver as pessoas utilizarem o Plumber para realizar coisas que eu nunca poderia ter feito.

### 11.2.5 Documentação

Quando uma API está funcionando, é um ótimo momento para se escrever a documentação dela. Quanto mais cedo no processo tiver a documentação, mais fácil será manter a API com o tempo. Na verdade, é uma boa prática escrever a documentação sobre sua API antes de escrever a primeira linha de código. Nesse caso, a documentação é seu blueprint para criar a API, e as pessoas que usarem seu modelo têm bastante tempo para se preparar para ela.

O núcleo da documentação da API é a especificação para as requisições da API: quais dados podem ser enviados para endpoints e o que se espera como retorno? Essa documentação permite que outras pessoas escrevam o código que chamará a API e saibam o que esperar. A documentação deve incluir muitos detalhes, como:

- As URLs de endpoint: (*<http://www.companywebsite.com/example>*).
- O que precisa ser incluído na requisição.
- O formato e o conteúdo da resposta.

A documentação poderia estar presente em qualquer documento de texto, mas também há modelos-padrão para armazená-la, como documentos OpenAPI. Um *documento OpenAPI* é uma especificação para escrever arquivos de especificação da API que podem ser facilmente compreendidos por usuários ou sistemas de computador.

O ideal é não ser a pessoa que mantém a API em execução para sempre, então é melhor ter a documentação sobre quais requisitos a API tem de

executar no seu sistema e como instalá-la em outro lugar. A documentação permite que a pessoa que assumir o código trabalhe por conta e faça alterações conforme necessário.

Por fim, é bom ter alguma documentação sobre o motivo pelo qual o modelo existe e os métodos básicos subjacentes a ele. Essa documentação é útil quando não estiver mais trabalhando no produto, e o conhecimento sobre por que foi criado não é perdido.

### 11.2.6 Teste

Antes de um modelo de machine learning ser colocado em produção e os clientes dependerem dele, é importante certificar-se de que funciona. Quando um modelo de machine learning é treinado, parte do processo é verificar o resultado do mesmo para assegurar a precisão, que é útil, mas não completamente suficiente para saber se funcionará. Em vez disso, o modelo precisa ser testado para verificar se pode lidar com qualquer entrada que possa receber sem falhar. Se um modelo de rotatividade que está sendo colocado em produção tiver uma API que usa uma ID numérica de cliente como entrada, o que acontece se a ID do cliente estiver em branco? Ou um número negativo? Ou a palavra *cliente*? Se, nesses casos, a API retorna algo que os usuários não esperam, isso pode ser ruim. Se a entrada incorreta causar a falha da API, pode ser catastrófico. Assim, quanto mais questões puderem ser percebidas antecipadamente, melhor.

Existem muitos tipos de testes, mas para criar um modelo de machine learning de produção, um tipo particularmente importante é o teste de unidades (unit testing) – o processo de testar cada componente pequeno do código para garantir que o sistema funcionará na prática. No caso de uma API de machine learning, muitas vezes significa testar se cada endpoint da API comporta-se como esperado em diferentes condições. O teste pode incluir receber como input (entrada) números muito grandes, números negativos ou strings com palavras estranhas. Cada cenário é transformado em um teste. Para um modelo de machine learning que classifica texto como um sentimento positivo ou negativo, o teste poderia ser “no input ‘eu te amo’, nós esperamos que a resposta da API seja positiva”. Outro teste pode ser que, se a entrada for um número como 27,5, o código retorna o

resultado “não foi possível calcular” em vez de falhar.

Além de testar os endpoints da API, também pode testar individualmente as funções dentro do código. O objetivo é ter 100% de cobertura, o que significa que cada linha de código na API foi testada para funcionar corretamente. Sempre que o modelo for implantado, os testes serão verificados e, se algum deles falhar, os problemas terão de ser resolvidos.

É fácil eliminar os testes, pois não há tempo para isso, mas muitas vezes é a única forma de detectar problemas importantes antes de o modelo ser apresentado aos clientes. Escrever uma grande quantidade de verificações parece um trabalho árduo comparado com a tarefa de construir um modelo de machine learning, mas é extremamente importante e não deve ser ignorado.

### **11.2.7 Como implantar uma API**

Se tiver um modelo de machine learning codificado para executar em seu computador, não é muito trabalhoso transformá-lo em uma API que seja executada nesse computador. Infelizmente, às vezes, você quer desligar seu computador ou utilizá-lo para assistir à Netflix; por isso, ter uma API sendo executada de forma contínua não é uma estratégia a longo prazo. Para que sua API seja executada de forma contínua e estável, ela precisa ficar em um servidor em algum lugar para que sempre seja executada. O processo de mover o código para ser executado em um servidor é o que queremos dizer com implantação. Configurar o servidor para que o código esteja sempre em execução dá um pouco mais de trabalho do que apenas criar a API.

## **O termo servidor**

Ouvir a palavra *servidor* pode ser intimidante, como se fossem computadores especiais que as pessoas normais não entendem. Na prática, um servidor é apenas um computador comum, como um portátil, mas que funciona em algum lugar sem uma tela. Em vez de ir até um servidor e fazer login, as pessoas conectam-se remotamente com outros computadores, que simulam estar na frente dele. Os servidores executam quase os mesmos sistemas operacionais que os computadores normais – Windows ou Linux com alguns pequenos ajustes. Caso se conecte remotamente a um servidor, ele deve lhe parecer muito familiar, com o mesmo menu Iniciar para Windows ou um terminal para Linux.

As vantagens dos servidores são que as pessoas geralmente os deixam em execução, em lugares mais seguros do que em home office. Mas não há razão alguma para não pegar um PC antigo, colocá-lo em seu armário e tratá-lo como um servidor; muitos que têm a engenharia de software como hobby fazem exatamente isso.

Quando as pessoas falam sobre usar serviços na nuvem, como Amazon Web Services (AWS), Microsoft Azure e Google Cloud Platform (GCP), querem dizer que estão usando servidores alugados da Amazon, Microsoft e Google. Mas só porque você está pagando a uma grande empresa por esses serviços não significa que os computadores sejam diferentes. Você pode pensar neles como computadores caros.

Há duas maneiras básicas de implantar uma API em um servidor: executá-la em uma máquina virtual ou colocá-la em um contêiner.

## **Como implantar uma API em uma máquina virtual**

Os *servidores corporativos* em geral são máquinas extremamente poderosas e caras. Raramente faz sentido ter uma dessas dedicadas a uma única tarefa, porque seria um exagero para a maioria das tarefas. Em vez disso, o servidor executará muitas tarefas ao mesmo tempo, mas seria catastrófico uma tarefa travar o computador e, então, fazer com que tarefas não relacionadas falhassem também. As máquinas virtuais constituem uma solução para esse problema porque são simulações de computadores. O computador grande e caro executará muitas simulações de outros computadores ao mesmo tempo. Se uma simulação falhar, não há problema; as outras continuam funcionando. Uma máquina virtual pode, em quase todos os casos, ser tratada exatamente da mesma forma que um computador normal; se fizer login em um computador, não poderá dizer que é uma máquina virtual, a menos que esteja procurando por ela. Toda vez que usar AWS, Azure ou GCP para acessar um computador, você está se conectando a uma máquina virtual. Se solicitar ao departamento de TI um servidor, provavelmente também farão uma máquina virtual para você.

As máquinas virtuais são ótimas porque são simulações: é possível ativá-las ou desativá-las facilmente. Você também pode obter snapshots para que possa voltar a uma versão anterior ou ter várias cópias em execução ao mesmo tempo. Você pode compartilhar um snapshot com outra pessoa, e elas também podem executá-lo. Ou pode fechar seus olhos e fingir que a máquina virtual é um computador velho normal e ficar bem com isso (supondo que usa seu computador com os olhos fechados).

Como uma máquina virtual é um computador normal, a maneira mais



simples de implantar seu código na máquina é instalar R ou Python, instalar as bibliotecas necessárias, copiar seu código para ele e executar seu código. Esses passos são os mesmos que usará para que sua API seja executada em seu computador! Se desejar fazer alterações na API, basta copiar a versão mais recente do código para a máquina virtual e executá-lo. Realmente é o caso de colocar um sistema em produção seguindo apenas estas etapas:

1. Inicie uma máquina virtual.
2. Instale os programas e o código necessários para executar a API do modelo de machine learning.
3. Comece executando sua API.

Levando em conta quantas pessoas falam sobre a complexidade de criar sistemas de produção, chama a atenção a facilidade com que pode fazê-lo de uma forma simples.

Um dos principais problemas com esse método simples de copiar e colar código em uma máquina virtual e pressionar Run (Executar) é a necessidade de mover o código manualmente sempre que fizer uma alteração. Esse processo é trabalhoso e propenso a erros. É fácil esquecer-se de mover o código ou perder o controle sobre qual versão está na máquina virtual.

A *integração contínua* (CI, em inglês) é a prática de ter o código recompilado automaticamente toda vez que ele recebe commit em um repositório. As ferramentas de CI podem monitorar repositórios git, ver quando as alterações são feitas e reconstruir o software com base nessa informação. Se estiver usando R ou Python, a recompilação provavelmente não será necessária, mas o processo de compilação pode executar etapas como a reexecução dos testes de unidade. A *implantação contínua* (CD, em inglês) é a prática de obter o resultado de ferramentas de integração contínua e implantá-la automaticamente em sistemas de produção. CI/CD referem-se ao uso de ambas as práticas.

Assim, uma ferramenta de CI/CD verificará seu repositório para detectar alterações e, se encontrar algumas, executará seu processo de compilação (como testes de verificação de unidade) e, em seguida, moverá o código resultante para uma máquina virtual. Como cientista de dados, não é preciso

se preocupar em fazer alterações na máquina virtual; a ferramenta CI/CD fará isso por você. Configurar uma ferramenta CI/CD por conta própria não é uma tarefa fácil, mas, se sua empresa tiver uma equipe de desenvolvimento de software, é provável que já tenha essas ferramentas configuradas e prontas para uso.

Outra melhoria no uso de máquinas virtuais é executar várias ao mesmo tempo. Se esperar que sua API obtenha muito tráfego, poderá fazer cópias da máquina virtual, executá-las todas ao mesmo tempo e atribuir tráfego aleatoriamente a uma máquina. Além disso, pode monitorar como cada máquina virtual está ativa e iniciar e parar cópias extras da máquina, conforme necessário. Essa técnica é chamada de *autoscaling*. Embora seja prático para sistemas grandes, autoscaling é um pouco difícil de configurar e, se estiver em uma situação na qual precisa dele, provavelmente também se encontrará em uma situação em que os desenvolvedores de software estão por perto para ajudá-lo.

## **Como implantar uma API em um contêiner Docker**

A configuração e a execução de uma máquina virtual podem ser um problema. Uma vez que cada uma é uma simulação de um computador, a configuração é tão irritante quanto a de um computador normal. É preciso instalar cada programa, alterar cada driver e configurar corretamente a máquina. É realmente difícil documentar todas as etapas necessárias e, se alguém repetir o processo, é fácil cometer erros. Outro problema é que, como cada máquina virtual é uma simulação de um computador, elas ocupam muito espaço porque têm de conter tudo o que um computador normal faz.

O Docker é uma solução para esses problemas. Para usar a metáfora de Mike Colemando blog do Docker (<https://www.docker.com/blog/containers-are-not-vms>), se um servidor cheio de máquinas virtuais é como um bairro de casas, os contêineres Docker são conjuntos de apartamentos em um único edifício. Embora cada apartamento seja uma unidade totalmente habitável, os apartamentos compartilham serviços como um aquecedor de água. Em comparação a máquinas virtuais, os contêineres Docker são muito mais fáceis de configurar e mais eficientes de executar.

O Docker permite especificar facilmente como uma máquina está configurada e, ao ter uma especificação compartilhada entre máquinas distintas, pode partilhar recursos. Isso permite que a criação e a manutenção de sistemas de produção sejam muito mais fáceis do que com máquinas virtuais, razão pela qual a Docker arrebatou o mundo do desenvolvimento de software.

Para entender o Docker, é importante compreender três conceitos:

- Um *dockerfile* é um arquivo de texto que contém todos os passos necessários para configurar a máquina simulada. Esses passos podem incluir “instalar Python 3” ou “copiar sobre o arquivo de modelo salvo na máquina”. A maioria das etapas é exatamente a mesma que os comandos bash do Linux; portanto, se estiver acostumado a ler esses comandos, um *dockerfile* deve parecer familiar.
- Uma imagem *Docker* é o resultado de quando o Docker segue os passos de um *dockerfile* para construir e armazenar um snapshot de um estado do computador.
- Um *contêiner* é criado quando o Docker pega uma imagem e começa a executá-la. Um contêiner em execução pode ser conectado e usado como um computador físico normal com os programas e dados especificados na imagem.

O Docker tem muitas vantagens em relação aos métodos tradicionais de implantação, mas o seu uso para implantar um modelo de machine learning na produção provavelmente funcionará apenas se outras pessoas na empresa estiverem usando contêineres Docker. Se for esse o caso, a empresa terá alguém que saiba como criar um contêiner Docker, implementá-lo e monitorá-lo para ver se está funcionando continuamente. Caso contrário, provavelmente haverá o retrocesso pelo fato de os modelos de machine learning terem sido implantados de forma não padronizada.

Como os contêineres Docker são mais complexos para começar do que as máquinas virtuais, se nunca tiver implantado código antes poderá ser mais fácil começar com máquinas virtuais. Mesmo que não seja possível utilizar o Docker para implantar modelos na produção, a sua utilização para análises reproduzíveis tem muitos benefícios. Vale a pena ganhar pelo

menos um pouco de experiência na utilização da ferramenta em algum momento, mesmo que não fique imediatamente claro de que é uma necessidade para sua carreira.

### **11.2.8 Teste de carregamento**

Se muitos sistemas estiverem usando seu modelo ou se um sistema estiver utilizando seu modelo muitas vezes ao mesmo tempo, é melhor certificar-se de que a API não falhará sob o estresse. Essa falha pode ocorrer porque o sistema em que a API está sendo executada está sem memória, porque o sistema leva muito tempo para processar cada requisição e tem uma fila cada vez maior ou por causa de qualquer outra coisa séria.

A maneira mais fácil de garantir que não aconteça é executar um *teste de carregamento* (load test) – um teste no qual faz um grande número de requisições à API ao mesmo tempo e se observa como a API se comporta. Normalmente, executa-se uma série de requisições que são pelo menos duas vezes mais do que poderia ser esperado. Se a API lidar com essas requisições com graciosidade, então deu tudo certo. No entanto, se falhar, saberá que precisa tornar o código mais eficiente, escalar seu sistema ou fazer outras alterações.

## **11.3 Como manter o sistema em funcionamento**

Mesmo depois que sua API for implantada e usada com êxito, não é suficiente. (Nunca é suficiente!) Você ou outra pessoa na empresa terá o trabalho de continuar a garantir que a API funcione. Algumas empresas têm uma equipe de operações de desenvolvimento (DevOps) cuja função é garantir que as APIs estejam sempre funcionando. Mesmo que a API esteja funcionando bem, ainda é possível precisar fazer ajustes por outros motivos. As seções a seguir discutem três considerações importantes para a manutenção da API ao longo do tempo.

### **11.3.1 Monitoramento do sistema**

É uma boa ideia seguir monitorando o funcionamento do modelo. Quantas requisições recebe a cada hora? As previsões do modelo ainda são precisas?

Há erros ocorrendo? A maneira mais fácil de controlar essas métricas é fazer com que sua API inclua log e telemetria. Fazer um *log* significa registrar dados sobre problemas internos na ferramenta, sempre que o modelo apresentar um erro, por exemplo. A *telemetria* é o registro de eventos que ocorrem, como, por exemplo, sempre que é feita uma requisição ou uma previsão específica. Além disso, o alerta pode ser configurado para quando ocorrerem problemas.

O log pode ser tão simples quanto ter suas informações de gravação de API em um arquivo todas as vezes que um evento ocorre. Em seguida, para verificar os logs, basta inserir o contêiner Docker ou a máquina virtual. A telemetria geralmente envolve o envio de informações de eventos para um local remoto (como um servidor centralizado), de modo que a telemetria de muitos sistemas esteja localizada em um local. Em seguida, é possível criar um dashboard para que a telemetria possa ser visualizada e monitorada em tempo real.

As ferramentas de alerta são usadas de modo que quando algo está indo mal, as pessoas na empresa descobrirão. Esses alertas podem ser emails automáticos ou mensagens no Slack enviadas quando ocorre um conjunto específico de eventos. Se a API do modelo tiver um evento de telemetria para quando uma requisição for recebida e nenhuma requisição ocorrer durante um dia inteiro, um email de alerta poderá ser enviado para informar que o sistema não está recebendo tráfego, mas provavelmente deveria estar recebendo.

Esses diferentes sistemas de monitoramento são frequentemente usados em conjunto, e as empresas tentam padronizar para que possam monitorar todas as APIs da empresa da mesma forma. Como em grande parte deste capítulo, quanto mais puder trabalhar com os padrões da sua empresa, mais útil será a sua ferramenta.

### **11.3.2 Como retreinar o modelo**

Muitas vezes, acontece que, em algum momento, depois de um modelo de machine learning ser colocado em produção, ele começará a não funcionar também. Os modelos de machine learning são treinados com base em dados e, à medida que o tempo passa, os dados tornam-se menos relevantes. Um

modelo de machine learning para prever rotatividade, por exemplo, pode falhar à medida que os clientes de novas regiões começam a se envolver com a empresa. Quando o modelo não tem desempenho suficiente, ele precisa ser retreinado.

A solução mais simples para o retreinamento é que quando o modelo vai mal, os passos que seguiu para treinar o modelo devem, em primeiro lugar, ser repetidos, mas carregando uma versão mais recente dos dados. Esse processo provavelmente significa carregar dados em R ou Python em seu computador, executar novamente os scripts e colocar o modelo em produção da mesma maneira. Essa é uma boa abordagem porque se fez algo uma vez, pode fazê-lo novamente. Na verdade, em grandes empresas, muitos sistemas importantes de machine learning de produção são tratados dessa forma.

Uma coisa que você pode fazer para tornar esse processo mais sofisticado é criar algum agendamento-padrão para o trabalho a ser feito. Em vez de tentar cuidar da métrica do modelo e usar seu instinto para decidir quando retreiná-lo, defina uma prática-padrão de fazê-lo a cada  $n$  semanas ou meses. Essa prática elimina as suposições de escolher quando realizar um ato importante.

E, mais importante, fazer o retreinamento em um horário-padrão significa que pode automatizar o processo. Se tiver um script em Python ou R que carrega os dados, constrói o modelo e o salva em algum lugar, é possível configurar um sistema para fazer essas coisas automaticamente conforme o agendamento definido. Na verdade, esse sistema de reaprendizagem pode ser colocado em produção por si só, para que não tenha de gastar tempo com isso. Esse sistema também pode testar se o modelo recentemente retreinado está funcionando da mesma maneira ou melhor do que o anterior; caso contrário, envie um alerta aos cientistas de dados. Processos de retreinamento assim avançados estão se tornando mais comuns, e as ferramentas de nuvem, como o AWS SageMaker, têm suporte para eles.

As pipelines de retreinamento automático são sofisticadas e estão em voga, mas, no fim do dia, desde que esteja retreinando seu modelo, está indo bem. Os cientistas de dados entram em apuros quando constroem um modelo, o implantam na produção e deixam de prestar atenção nele à

medida que o modelo vai piorando ao longo do tempo. Ao não continuar a monitorar o desempenho de um modelo e repará-lo, se necessário, corre o risco real de danificar seu trabalho em vez de ajudar. Fique atento!

### **11.3.3 Fazer alterações**

Se seu modelo de produção for bem-sucedido para a empresa, inevitavelmente irá querer fazer alterações no modelo para melhorá-lo. Talvez queira obter mais conjuntos de dados ou alterar o método de machine learning para melhorar o desempenho da API, por exemplo. Também ouvirá pessoas da empresa sobre os recursos que desejam no modelo ou nos problemas que eles encontram com ele.

Como na discussão de análises no Capítulo 10, o tipo de alterações “mais uma coisinha” pode apresentar questões reais. Não fica claro se fazer esse trabalho vale necessariamente o tempo investido, mesmo que seja interessante ou pareça importante para alguém. Se levar três meses para o modelo passar de 84% a 86% de precisão, terá perdido três meses que poderiam ter sido gastos em outro projeto. Ou um recurso que parece importante para um stakeholder em particular, pode não afetar muitos clientes. Um modelo de machine learning bem-sucedido na produção chamará a atenção de muitas pessoas e, como o cientista de dados que ajudou a criá-lo, deve tentar garantir que o tempo investido em melhorá-lo seja bem aproveitado.

## **11.4 Encerramento**

Este capítulo aborda muitos conceitos sobre a implantação de modelos, alguns dos quais podem ser familiares ou não. Embora nem todos os tópicos possam ser relevantes para seu trabalho, é ótimo ter uma noção básica caso venha a precisar deles no futuro. É possível encontrar muitos recursos bons em livros e online que oferecem mais informações sobre esses tópicos, especialmente porque há muita sobreposição com a engenharia de software. Conforme a ciência de dados segue mudando como área, esses tópicos continuarão sendo importantes e vale a pena continuar a aprender.

## **11.5 Entrevista com Heather Nolis, engenheira de machine learning da T-Mobile**

Heather Nolis, mestre em ciência da computação e bacharel em neurociências e em francês, é engenheira de machine learning da equipe de IA na T-Mobile, onde ajuda a colocar em produção modelos em R e em Python que são clicados milhões de vezes por semana.

### **O que significa “engenheiro de machine learning” na sua equipe?**

Eu pego os modelos que os cientistas de dados fazem e os projeto em produtos que a equipe mantém. Durante muito tempo, a T-Mobile tinha cientistas de dados que ficavam construindo modelos bonitos e conduzindo análises bacanas que depois enviariam ao departamento de engenharia de software para colocar em produção. A ideia era que isso deixaria o trabalho ter um impacto real na empresa, mas era difícil para os engenheiros utilizarem o trabalho, pois havia uma enorme barreira linguística entre cientistas e engenheiros de dados. Meu objetivo é fazer o meio-campo para entender todas as coisas que entram em uma análise ou que são importantes para um determinado modelo e comunicá-las aos engenheiros.

### **Como foi implantar seu primeiro código?**

Minha primeira implantação foi na primeira semana que atuei como desenvolvedora de software. Foi num produto já existente e, inicialmente, não entendi por que era arriscado. Quando se programa em um computador, é normal executar o código 50 vezes para testá-lo. Mas na fase de produção, se uma parte do código tiver um erro, pode causar um enorme problema para a empresa, pois não se trata só de você no seu computador ficando incomodado por executar um código que não funciona. Na minha versão inicial, quando pensei que estava pronto, tive de fazer três horas de teste de integração antes de poder liberá-lo.

### **Se há problemas na produção, o que acontece?**

No início, construí uma ferramenta baseada no Twitter que recomendaria a



loja T-Mobile mais próxima nas redes sociais. Escrevi o código em Node.js, que a equipe não tinha suporte, mas pensei: “Vou resolvê-lo para mim e mostrar às pessoas que poderia ser feito e alguém mais qualificado pode fazê-lo”. Foi aí que aprendi que “alguém mais qualificado pode fazê-lo” nunca acontece; foi o meu código que foi colocado na produção.

Liberamos o código e, sendo totalmente nova no Node, não foi um código bonito. Funcionou e era seguro, mas, tendo uma experiência de produção limitada, faltava confiança de que funcionaria na produção. Outros engenheiros estavam nervosos porque era uma linguagem que ainda não tínhamos suporte. Pelos próximos dois meses, fui chamada todas as vezes quando um serviço tinha um problema qualquer. Tinha de estar lá porque arrisquei liberar algo em uma linguagem diferente e em uma plataforma nova; as pessoas imaginavam que tudo que era estranho vinha de mim.

Cada vez que era notificada naqueles dois meses inteiros, sentia como se fosse um problema meu. Mas nunca era o meu código que quebrava na produção! Penso que é algo a ser lembrado: ao colocar coisas em produção pela primeira vez, há a chance de quebrar tudo, mas as coisas de outras pessoas também não funcionam. Não é apenas o seu código. Não precisa ter medo quando as coisas se quebram!

Também digo isso sobre colocar as coisas em produção: é claro que gostaria sempre de construir os modelos mais bonitos, mas nem sempre ajuda construir o produto que queremos. No fim, temos de sacrificar muitas dessas coisas para termos um código robusto que continuará funcionando. Meu trabalho como engenheira de machine learning é compreender essas compensações e impulsionar a criação de um produto.

## **Qual é seu último conselho para cientistas de dados que trabalham com engenheiros?**

Dois pontos importantes para trabalharem bem juntos são: compreender a linguagem deles e valorizar aquilo com que se importam. Para entender a linguagem deles, considere coisas que podem parecer uma frase normal para você, mas que soa a um engenheiro de machine learning como entrar em casa e encontrar uma TV antiga de tubo. Eles se perguntam: “Eu voltei no tempo?”. Meu exemplo favorito foi quando recebemos nosso primeiro

cientista de dados na equipe de IA na T-Mobile. Em um momento perguntei a ele: “Não podemos simplesmente colocar o modelo em R na produção como uma API?”. Ele me perguntou: “Você quer que eu execute R como um servidor web?”. Voltei no tempo por um segundo porque quando escuto as palavras *servidor web*, na verdade eu escuto: “Olá, estou vindo da década de 1980!”. Fiquei bem desanimada, mesmo que significasse exatamente a mesma coisa.

Para o último ponto, no fim do dia, os cientistas de dados sentem-se bem em relação ao trabalho deles quando estão criando modelos precisos. Como engenheira, o que me faz me sentir bem é colocar coisas em produção para que outras pessoas possam tocá-las. A única coisa que realmente valorizo é o código de trabalho. Se puder chegar em um engenheiro e dizer: “Projetei uma API e criei um documento que especifica todas as entradas e saídas”, isso mostra aos engenheiros que você está pensando nos problemas que eles também enfrentam.

## Resumo

- Implantar na produção é a prática de fazer com que os modelos sejam executados de forma contínua.
- Colocar um modelo em uma API REST permite que outros sistemas a utilizem.
- As APIs podem ser implantadas em máquinas virtuais ou como contêineres Docker.
- Examine bem como sua empresa gerencia o código, os testes e a implantação de sistemas de produção.

# CAPÍTULO 12

## Como trabalhar com stakeholders

Este capítulo abrange:

- Como trabalhar com diferentes tipos de stakeholders
- Como lidar com pessoas fora da equipe de ciência de dados
- Como ouvir melhor para que seu trabalho seja mais bem utilizado

Parece que o trabalho de um cientista de dados seria principalmente sobre dados, mas na sua grande parte gira em torno das pessoas. Os cientistas de dados passam horas ouvindo as pessoas na empresa falarem sobre os problemas que têm e como os dados podem resolvê-los. Os cientistas de dados têm de apresentar o trabalho às pessoas para que elas possam utilizar os conhecimentos adquiridos por meio de uma análise ou confiar em um modelo de machine learning. Quando ocorrem problemas, tal como projetos atrasados ou dados indisponíveis, é preciso conversar com essas pessoas para saber qual deve ser o próximo passo.

Karl Weigers e Joy Betty definem o termo stakeholder (partes interessadas) em *Software Requirements* como “uma pessoa, um grupo ou uma empresa que esteja ativamente envolvida em um projeto, é afetada pelo resultado ou pode influenciar no resultado”. Para um cientista de dados, os stakeholders podem ser pessoas que trabalham no marketing, desenvolvimento de produtos ou outras áreas da empresa que usam a ciência de dados para tomar decisões. Também podem ser pessoas da engenharia que dependem de modelos de machine learning criados por cientistas de dados para alimentar o software delas ou garantir que os dados sejam coletados de maneira adequada. Em algumas situações, são executivos de alto nível. Os stakeholders podem vir de toda a empresa, e diferentes stakeholders têm comportamentos e necessidades distintos.

Neste capítulo, explicamos o que esperar dos diferentes tipos de

stakeholders que encontrará em um projeto de ciência de dados. Em seguida, abordamos como trabalhar com eles de forma eficaz e como deve elaborar a comunicação com pessoas fora da equipe de ciência de dados. Por último, apresentamos o processo de priorização do trabalho que os stakeholders lhe dão.

## 12.1 Tipos de stakeholders

Cada stakeholder que encontrar em um projeto de ciência de dados tem sua própria experiência e motivação. Embora um stakeholder possa ser qualquer pessoa, dependendo de quão estranho seu projeto for, a maioria dos stakeholders recai em uma de quatro categorias: administração, engenharia, liderança e seu gerente (Figura 12.1).



*Figura 12.1 – Tipos de stakeholders abordados nesta seção.*

### 12.1.1 Stakeholders da administração

Os *stakeholders da administração* são pessoas de um departamento, como marketing, atendimento ao cliente ou produto, que supervisionam as decisões da empresa, aquelas que solicitam análises para ajudá-las a tomar melhores decisões ou modelos de machine learning para aumentar a eficiência. As pessoas dessas funções têm experiências variadas: alguém no marketing poderia ter um MBA e ter vindo de uma agência de publicidade, embora um gerente no departamento de atendimento ao cliente poderia ter começado como um agente de atendimento com nível técnico e ter subido de posto. Esses caminhos variados para o trabalho dão a cada pessoa uma perspectiva diferente ao trabalhar com você.

Geralmente, os stakeholders da administração têm pouca formação técnica. Podem ser competentes em Microsoft Excel, mas, com algumas exceções, é o máximo de experiência analítica. A maioria dos stakeholders da administração não sabe como usar R ou Python, nem os méritos de diferentes modelos de machine learning. Mas se tiverem lido algum artigo na última década, ouviram repetidamente falar sobre o valor dos dados na tomada de decisões e sobre a importância da ciência de dados. Assim, os stakeholders da administração estão em uma situação delicada: têm de confiar nos cientistas de dados para que lhes forneçam as informações cruciais de que necessitam para tomar decisões ou lançarem ferramentas de machine learning. Todavia, sem os conhecimentos técnicos, precisam confiar que aquilo que o cientista de dados diz é correto.

Com frequência, um stakeholder da administração está altamente envolvido em um projeto de ciência de dados. Eles estão ali para ajudar a dar o pontapé inicial e definir o objetivo do projeto. Ficam ali durante o projeto para examinar os resultados intermediários e dar feedback para a empresa. E estarão lá na conclusão do projeto, quando a análise final é apresentada ou o modelo é implantado. Como são aqueles que garantem que a empresa é valorizada com o trabalho de ciência de dados, eles têm de estar constantemente envolvidos.

Como cientista de dados, é seu trabalho entregar-lhes o que eles precisam para que a parte deles funcione, como uma análise, um dashboard ou (ocasionalmente) um modelo de machine learning. Não só precisa fazer esse trabalho para eles, mas também precisa ter certeza de que eles entendem e confiam nele. Se lhes der uma tabela com estatísticas complexas e nenhuma explicação, eles não a entenderão e, portanto, não poderão usá-la. Sendo um parceiro de negócios confiável para eles, permite-se assim que eles usem os dados e oferecem oportunidades para que mais ciência de dados seja feita na empresa.

As situações mais difíceis com os stakeholders da administração tendem a ser quando eles não aceitam os resultados da ciência de dados, como quando um cientista de dados faz uma análise e o stakeholder responde: “Ah, mas não pode estar certo”. Quando os fatos e as premissas do trabalho de ciência de dados são questionados, há a possibilidade de o stakeholder

tirar os cientistas de dados da jogada. Nessas situações, o melhor a fazer é ajudá-los a entender o que fez e como fez. Geralmente, a falta de confiança vem da não compreensão, mas, conversando sobre como as coisas foram feitas, talvez seja necessário mudar as suposições na análise.

### **12.1.2 Stakeholders da engenharia**

As equipes de engenharia são responsáveis pela manutenção do código (e, potencialmente, pelos produtos físicos) que a empresa oferece, e, quando esses produtos requerem algoritmos de machine learning ou análises de ciência de dados, elas tornam-se stakeholders. Em alguns aspectos, é mais fácil se trabalhar com os engenheiros do que com outros tipos de stakeholders, porque, culturalmente, têm muitas semelhanças com os cientistas de dados. Tal como os cientistas de dados, eles têm uma formação técnica que receberam de universidades, bootcamps ou cursos online.

Embora os engenheiros tenham uma vasta formação técnica, muitas vezes têm pouca experiência com os componentes cruciais do trabalho de ciência de dados. Embora um engenheiro de software programe, geralmente o faz com uma tarefa extremamente específica em mente, como criar uma API que consulta um banco de dados específico. O trabalho de um desenvolvedor de software não tem o componente exploratório do trabalho de um cientista de dados, portanto é estranha a ideia de passar semanas tentando entender dados.

Os engenheiros tendem a colaborar com cientistas de dados quando um modelo de machine learning é necessário como parte de um projeto de engenharia. Essa colaboração é mais frequentemente realizada como um modelo de machine learning sendo convertido em uma API em produção que os engenheiros utilizarão para o trabalho que desenvolvem (consulte o Capítulo 11). Os engenheiros dependem dos cientistas de dados e dos engenheiros de machine learning para criarem um produto que tenha entradas e saídas claras, que seja confiável de utilizar e que não os surpreenda na produção. Como cientista de dados, é seu trabalho apresentar esses detalhes a um stakeholder da engenharia. Você tem de pensar como um engenheiro e tentar entender qual seria o melhor produto para suas necessidades.

Os engenheiros também dependem de cientistas de dados para fazer análises que ajudem a alimentar as ferramentas que estão sendo construídas. Os cientistas de dados podem analisar os dados para ajudar a priorizar recursos, diagnosticar bugs nos sistemas de engenharia e avaliar o desempenho de produtos voltados ao cliente, como sites. Nessas situações, os engenheiros estão mais próximos dos stakeholders da administração, pois precisam que os cientistas de dados obtenham para eles o conhecimento necessário para tomar as decisões certas.

As dificuldades tendem a surgir com os stakeholders da engenharia em torno da incerteza do trabalho em ciências de dados. Ao desenvolver um produto de software, geralmente você projeta uma API ou um processo, e acaba desenvolvendo. As tarefas e os requisitos do projeto são claros do que deve ser feito. Na ciência de dados, por outro lado, há poucas expectativas na criação de um produto. Não está claro quais dados precisará, porque talvez não se saiba o que seria importante para o modelo. Não está claro, também, qual será o resultado, porque depende muitas vezes do seu modelo e desempenho. Sequer fica claro se a ideia será viável, porque talvez perceba que nenhum modelo é preciso o suficiente para atender às expectativas da empresa.

Como os projetos de ciência de dados são desconhecidos no início, os stakeholders da engenharia ficam muitas vezes surpresos pela pouca quantidade de dados que os cientistas podem prometer no início. Assim, como cientista de dados, precisará tomar cuidado extra para comunicar o processo para que os engenheiros sejam menos surpreendidos quando as coisas mudam. Certifique-se de comunicar com antecedência e frequência qual é o processo de ciência de dados e como está seguindo nele. Quando comunicar as ambiguidades da ciência de dados aos engenheiros, terão menos probabilidade de ficarem surpresos com eles.

### **12.1.3 Liderança corporativa**

Os executivos de uma empresa têm formações similares aos dos stakeholders (ou engenheiros, se liderarem uma empresa de tecnologia), mas esferas de influência muito maiores. Esses diretores, vice-presidentes e diretores corporativos estão liderando a empresa e precisam de dados para

isso. Os cientistas de dados são frequentemente encarregados de criar elementos para apresentar uma concepção daquilo que os executivos precisam para fazer os trabalhos deles. Os cientistas de dados também podem ser responsáveis diante deles se estiverem envolvidos em um projeto em grande escala, do qual machine learning é um componente crucial.

Os líderes corporativos são extremamente ocupados e têm pouco tempo para compreender detalhes que não os afetam. Esse fato leva a uma imensa dificuldade a conseguir um momento com eles, e, quando esse tempo é concedido, ele é curto. Quando se encontra com um executivo de alto nível, geralmente querem chegar ao ponto e entender de imediato as implicações. Faz sentido: são pessoas extremamente ocupadas e, quanto menos trabalho tiverem que fazer para descobrir o que alguém está tentando dizer, mais podem se dedicar à tomada de decisões.

Os líderes corporativos tendem a trabalhar com cientistas de dados quando precisam de dados para tomar uma decisão importante ou quando querem ter uma compreensão melhor de uma parte da empresa. Às vezes, o trabalho que é feito para stakeholders da administração ou para outras pessoas é apresentado a alguém de nível mais alto na empresa, inclusive para o primeiro escalão. Nessas situações, a análise e o relatório podem ser reorganizados muitas vezes à medida que vai subindo na escala. Em outras ocasiões, um executivo pode solicitar que uma análise ou trabalho específico seja feito, e é tarefa do cientista de dados criar algo para alguém que vê os resultados pela primeira vez e os entenda de imediato.

Dependendo do tamanho e da cultura da empresa, os cientistas de dados podem ter de trabalhar muito antes de compartilhar os resultados com um executivo. Em algumas organizações, equipes revisam os resultados para garantir que estejam alinhados aos objetivos e crenças da empresa. Naquelas mais tranquilas ou pequenas, os cientistas de dados podem criar o trabalho diretamente para o líder. Independentemente da empresa, o trabalho deve estar sempre bem feito e sem erros.

O problema tende a ocorrer quando o trabalho apresentado não está tão claro ou se encontra incompleto. Se o executivo não puder entender o que está sendo apresentado, não terá paciência para esperar que as coisas fiquem claras por conta própria. Se fizerem perguntas que o cientista de



dados não consegue responder, eles podem sentir que o trabalho não é confiável. Se a colaboração com o executivo não tiver êxito e o mesmo mostrar-se insensível, pode ser um forte golpe para a equipe.

Por outro lado, se o executivo gostar dos resultados ou achar que têm valor, pode ser ótimo para os cientistas de dados. Ao se tornar um parceiro confiável para um executivo, uma equipe de ciência de dados pode ganhar vantagens em uma empresa para usar dados e machine learning em mais lugares e por mais motivos.

### **12.1.4 Seu gerente**

Dependendo do projeto em mãos, seu gerente é às vezes um stakeholder. Se ele atribuir-lhe uma tarefa, sempre peça confirmação e faça sugestões, pois ele é efetivamente um stakeholder no projeto. Um gerente quer que um projeto seja bem-sucedido porque (1) o trabalho dele é fazer com que você também seja bem-sucedido, (2) fica bom para ele se os projetos atribuídos à equipe saírem-se bem e (3) o projeto pode se alinhar com os objetivos mais amplos do seu gerente para a equipe.

Em geral, seu gerente deve orientá-lo no projeto. Quando encontrar dificuldades, é preciso ser capaz de falar com ele sobre os problemas, que deverá ajudá-lo a encontrar o melhor caminho. Seu gerente pegará seu trabalho e irá auxiliá-lo a ir o mais longe possível na empresa, contando às pessoas sobre isso, ajudando a integrar o trabalho nos processos existentes e pensando em novas oportunidades para melhorar seu projeto.

Mas um gerente também é um stakeholder porque está confiando em você para fazer o trabalho. Ele precisa que você faça o melhor trabalho possível, pois seus relatórios, modelos e análises são o que o seu gerente compartilha. Por isso, um gerente desempenha uma função dupla como alguém em que se pode confiar para receber ajuda e como alguém com quem contar para fazer o trabalho.

Por causa dessa dupla função, o restante deste capítulo está relacionado ao seu gerente. A principal diferença é que, com o seu gerente, você pode relaxar e mostrar uma vulnerabilidade maior. É sensato dizer algo como: “Estou tendo muita dificuldade para concluir esta análise” ao seu gerente, mas provavelmente não diria isso a um executivo. Um gerente é capaz de

tratá-lo de maneira mais humana e lhe dar conselhos, enquanto os outros stakeholders são puramente clientes do seu trabalho.

No geral, trate os gerentes da mesma forma como trataria outras pessoas com quem trabalha: dê-lhes atualizações claras, comunique-se sempre e faça um trabalho apresentável. Mas, quando estiver com problemas, fale primeiro com seu gerente, porque ele quer ajudá-lo e saber quando as coisas precisam de assistência.

## **12.2 Como trabalhar com stakeholders**

Para se comunicar de forma eficaz com os stakeholders durante seus projetos de ciência de dados, existem quatro princípios fundamentais que devem ser considerados:

- Compreender os objetivos dos stakeholders.
- Comunicar-se com frequência.
- Ser consistente.
- Construir um relacionamento.

As seções a seguir abordam em detalhes cada um desses princípios.

### **12.2.1 Como entender os objetivos dos stakeholders**

Todos têm objetivos ao fazer um trabalho, os quais pretendem alcançar quando vão diariamente trabalhar. Esses objetivos são determinados pelo trabalho que uma pessoa tem e seus traços pessoais, como a ambição e um desejo de equilíbrio entre vida e carreira. Um engenheiro-chefe, por exemplo, pode concentrar-se na conclusão de um projeto para receber uma promoção. Ou um executivo sênior sabe que irá embora da empresa logo e não quer abandonar o barco subitamente. Esses objetivos moldam o que as pessoas fazem no trabalho e a forma como respondem às ações de outras pessoas. Um projeto que leva mais tempo do que o esperado pode ser terrível para um engenheiro voltado para a promoção, mas ótimo para o executivo que não quer fazer nada.

Sendo você um cientista que está trabalhando com stakeholders, é fundamental compreender suas metas. Exatamente a mesma análise poderia

ser bem recebida ou não, dependendo do ponto de vista do stakeholder. Considere uma análise do desempenho de determinado produto que está sendo vendido no site da empresa. Suponha que o stakeholder é a pessoa que controla esse produto, e, em sua análise, percebeu-se que o produto não estava vendendo bem na América do Sul. Se o stakeholder tivesse o objetivo de fazer o produto parecer excelente à empresa porque foi ideia dele, a análise poderia ser muito mal recebida, por destacar o problema. Por outro lado, se o stakeholder tivesse o objetivo de manter um portfólio inteiro de produtos com bom desempenho, saber qual deles pode ser cortado é muito útil.

Ao trabalhar com um stakeholder, tente compreender suas metas e motivação o mais rapidamente possível. Quanto mais rápido puder entender isso, será menos provável que você apresente algo que será desnecessariamente mal-recebido. Há algumas maneiras de descobrir as motivações de uma pessoa:

- *Pergunte a eles diretamente.* Ao perguntar a um stakeholder: “O que é importante para você?”, está lhe dando uma abertura para que ele se revele. O que as pessoas dizem diretamente não é a visão completa, mas, muitas vezes, é possível captar os elementos essenciais dessa forma. Além disso, é uma pergunta totalmente normal que deve ser feita em uma reunião introdutória.
- *Pergunte ao redor.* Pergunte se seus colegas de trabalho já trabalharam com os stakeholders antes. Fazer uma pergunta a alguém da equipe como: “Então, fale-me sobre esse stakeholder: como ele é?” pode fazer com que seus colegas de trabalho falem sobre ele. Evite fofocas; não espalhe o que um colega lhe disse em confiança.
- *Deduz a motivação dos stakeholders conforme as ações deles.* Às vezes, com base no que o stakeholder está fazendo, sua motivação pode ficar evidente. Se você estiver apresentando uma análise que reflita negativamente sobre um dos produtos que ele gerencia, por exemplo, e ele ficar excessivamente defensivo, é um sinal de que o produto é muitíssimo importante para ele. A desvantagem desse método é que você precisa aprender por meio da interação, e pode facilmente cometer erros, mas, se for inevitável interagir com ele, você também pode passar

um tempo aprendendo com esses erros.

Ao participar dessas tarefas, construa um modelo mental do stakeholder. Como reagirão a diferentes resultados de uma análise ou atrasos em um modelo? Se puder pensar nos resultados com antecedência, é possível ser cauteloso na comunicação.

Note que compreender as motivações dos stakeholders não significa ter de atendê-las. Embora a compreensão dos objetivos dos mesmos o ajude a prever a forma como reagirão, pode haver casos nos quais seus objetivos não se alinham com os dos stakeholders, tendo que ignorar os objetivos deles. Se seu objetivo é ser o melhor cientista de dados possível, na situação do exemplo em que sua análise demonstra que um produto vai mal, seria do seu interesse ser honesto em sua análise e não esconder seus achados. O conhecimento das necessidades dos stakeholders é algo que pode ajudá-lo.

Se tiver que informar algo que o stakeholder não gostará, é melhor chamar reforços. Seu gerente ou alguém mais experiente da equipe pode ajudar? Ao ter alguém que o auxilie a comunicar essa mensagem, eles serão capazes de navegar por consequências ou problemas políticos em vez de você. Não se espera que um cientista de dados iniciante seja um especialista na política da empresa e no panorama geral.

Quando tiver que ter uma conversa difícil por conta própria, é melhor tentar encará-la como uma colaboração. Pense que você está no mesmo lado que o stakeholder. O comunicado pode ser difícil de ouvir, mas dá para tentar convencê-los de que não está intencionalmente causando decepção, mas tentando ver a situação pela perspectiva deles e buscando oportunidade para superar a questão atual. Nesse caso, a conversa é muito mais uma negociação e uma discussão de negócios do que algo técnico; trata-se de ter diferentes perspectivas e chegar a um entendimento comum.

## **Indicadores-chave de desempenho (KPIs – Key Performance Indicators)**

Os indicadores-chave de desempenho (KPIs, do inglês) e os resultados-chave de objetivo (OKRs) são métricas nas quais uma equipe ou empresa se concentra, pois orientam o valor corporativo. Uma equipe de varejo online, por exemplo, pode concentrar-se em pedidos mensais como um número que pretende aumentar. Os KPIs são úteis para os cientistas de dados porque fornecem quantificação explícita dos objetivos da equipe. Se conseguir descobrir os KPIs de uma equipe, poderá enquadrar todas as suas análises e outros trabalhos em termos de como estes afetam os KPIs. Se uma análise ou método não estiver relacionado a um KPI, a equipe provavelmente não estará interessada.

Nem todas as equipes têm os KPIs principais e, por vezes, estão constantemente mudando ou são mal definidos, mas, se receber dados de KPIs, é melhor não ignorá-los. Muitas vezes, são a forma mais fácil de compreender rapidamente os objetivos do stakeholder.

### **12.2.2 Como comunicar constantemente**

É fácil para um cientista de dados ter preocupação a respeito de estar se comunicando demais ou de menos. “Enviar um email a um stakeholder pela terceira vez em um dia é muito?” é um pensamento recorrente ao clicar “Enviar mais uma vez”. Ou, com a mesma facilidade, poderia pensar: “Não falo com nosso stakeholder há algum tempo. O que será que está pensando?”. Ou o pior caso: talvez nem leve em consideração o quanto você deveria estar informando o stakeholder, fazendo com que ele desconheça completamente o andamento do projeto.

Para um cientista de dados, quase sempre é o caso de não estarem se comunicando o suficiente. Emails, reuniões e ligações são as únicas maneiras pelas quais os stakeholders podem entender o que está acontecendo em um projeto. Sem uma comunicação adequada, os stakeholders podem acreditar que não estão sendo informados e se preocuparem com o fato de não terem uma ideia correta do que está acontecendo. Todavia, se a comunicação for insuficiente, quando um cientista de dados falar com um stakeholder, ele pode se abalar com a diferença entre as expectativas e a realidade.

Um cientista de dados deve enviar várias mensagens aos stakeholders:

- Um cientista de dados deve manter o stakeholder informado sobre o cronograma do projeto. Se, no início do projeto, parecia que iria levar um mês para encontrar e limpar os dados e igual tempo para construir um modelo, pergunte ao stakeholder se esse cronograma ainda é o esperado. Se possível, o cientista de dados deve compartilhar mudanças e atrasos à medida que acontecem. Um cenário mostra-se adverso quando o stakeholder espera que um projeto tenha terminado, mas o cientista de dados ainda tem semanas ou meses de trabalho e não o informou sobre isso. Quando o stakeholder finalmente descobrir, pode ficar incomodado, e com razão.
- Um cientista de dados deve informar o andamento do projeto, como, por exemplo, os achados que o cientista de dados descobriu durante o projeto ou áreas nas quais está encontrando mais dificuldade do que o esperado. Algo como ficar bloqueado por causa de acesso a um banco de dados pode ser potencialmente resolvido com o auxílio do stakeholder. Compartilhar onde a análise está indo bem pode ajudar o stakeholder a melhorar o escopo do projeto. Se o projeto parece estar indo mal, também deve ser comunicado. (O Capítulo 13 fornece mais informações sobre projetos que falham.)
- Diferentemente de como o projeto está progredindo, um cientista de dados deve sempre atualizar o stakeholder sobre como o trabalho traz informações à empresa e o que vem a seguir. Deve ter opinião se o que foi feito até o momento deve mudar o direcionamento do projeto. Se, por exemplo, um cientista de dados estiver realizando uma análise e encontrar algo totalmente novo, deve criar um conjunto de recomendações para a empresa sobre o que fazer com essa informação.

Muitas vezes, a melhor maneira de estabelecer essa comunicação consistente é torná-la o padrão de como o projeto funciona. Nada é melhor do que uma reunião recorrente no calendário entre o cientista de dados e o stakeholder. Com uma reunião semanal ou quinzenal, garante-se uma comunicação básica. Essa rotina tem uma função obrigatória: ao tê-la no calendário, você se obriga a fazer algo para compartilhar a cada reunião. Para cada uma dessas reuniões, você deve vir munido com uma lista de atualizações do cronograma, notas sobre o que está indo bem ou mal, partes

do trabalho para compartilhar e sugestões para as etapas seguintes.

Como cientista de dados, tenha também o hábito de enviar emails diretamente aos stakeholders, conforme necessário. Como aspirante ou cientista de dados iniciante, talvez acredite ser muito intimidador fazer perguntas a pessoas mais experientes na empresa. Na maior parte do tempo, porém, os stakeholders ficarão satisfeitos em responder a perguntas se significarem que seu trabalho será melhor; esse é o papel deles na empresa. Se estiver preocupado com o fato de a pessoa ser do alto escalão e que o email precisa ser superformal, ou se acha que as perguntas parecem óbvias o suficiente para pensarem que você não sabe muito, mostre primeiro o email ao seu gerente, pois essa é a função do seu gerente. Dependendo do stakeholder e do projeto, envie emails uma vez por semana mais ou menos.

Se as coisas mudarem subitamente em seu projeto (talvez um conjunto de dados que achou que não existia) e precisar de informações dos stakeholders, às vezes, o caminho certo é fazer com eles uma chamada ou reunião improvisada. Essas reuniões rápidas podem ser excelentes para receber retorno imediato quando necessário. A única pergunta a fazer a si mesmo antes de agendar uma dessas reuniões é: “Realmente preciso da contribuição do stakeholder sobre essa questão?”. Se ocorrer uma mudança, mas souber o que fazer a seguir, provavelmente não precisará tomar o tempo dos outros. Se precisar do retorno, siga em frente. Um erro que os cientistas de dados iniciantes tendem a cometer é assumir ser padrão que outras pessoas marcam reuniões, não eles. Mas quanto mais proativo for sobre manter o projeto andando, melhor o projeto se desempenhará. Além disso, tal situação é um excelente treinamento para funções mais especializadas nas quais essas ações são esperadas.

O método e os motivos da comunicação devem variar com base no tipo de participante. Em geral, os stakeholders da empresa tendem a ficar satisfeitos com reuniões nas quais podem orientar e colaborar. Provavelmente não são o tipo de pessoa que fornecem dados reais ou ajudam com questões técnicas. Os engenheiros frequentemente têm as respostas técnicas, mas serão tão incertas quanto você ao tomar decisões sobre o projeto ou a direção do trabalho. Os executivos são extremamente ocupados e em geral são informados somente no início do projeto para definir objetivos amplos e

no fim do projeto para ver a conclusão.

### 12.2.3 Como ser consistente

Imagine um restaurante na sua rua. Um dia, você encomenda fajitas ali, e eles lhe trazem as melhores fajitas que já comeu. Um mês depois, encomenda fajitas novamente, mas, desta vez, esqueceram-se totalmente de temperar a carne. Em uma terceira vez, a comida estava saborosa, mas levou mais de uma hora para ficar pronta – muito mais do que esperava. É um restaurante onde gostaria de comer?

As empresas prosperam com o fornecimento de um produto consistente e, como cientista de dados, você é uma miniempresa dentro da sua empresa. Os stakeholders são seus clientes e, se não servi-los bem, eles não irão mais lhe pedir ajuda. Uma maneira de oferecer consistência em suas relações é por meio da padronização do trabalho.

No caso de análises e relatórios, é possível contribuir muito com os stakeholders criando uma estrutura consistente para passar a informação. Se puder manter as coisas da mesma maneira a cada análise, os stakeholders poderão focar nos achados. A seguir, alguns aspectos a considerar na padronização:

- *Como a análise é estruturada* – tanto quanto possível, tente ter um formato para a análise. Comece com o mesmo estilo de Objetivos e Dados, terminando com conclusões semelhantes e próximas etapas. Assim treinará os stakeholders a lerem e analisarem esses materiais.
- *Como a análise é apresentada* – embora não tenha que seguir à risca, as coisas tendem a andar melhor se tiver um tipo de arquivo para suas análises. Esses arquivos podem ser em PowerPoint, PDF, HTML ou em outro formato. Todos devem ser sempre armazenados no mesmo local. É possível criar uma pasta no Dropbox, na rede ou em outra ferramenta de compartilhamento para suas análises. Certifique-se de que os stakeholders conseguem usar a ferramenta; um repositório GitHub provavelmente não funcionará, embora possa fazer um para si mesmo a fim de manter tudo sob o controle de versão.
- *Como a análise está desenhada* – esse ponto pode parecer insignificante,



mas a consistência nos visuais pode ir muito longe. Use as mesmas cores e modelos tanto quanto possível (pontos de bônus para as cores da sua empresa).

Quando estiver apresentando dashboards, muitas das regras de consistência para análises aplicam-se aqui. É melhor manter o estilo e o formato consistentes entre os vários dashboards e armazená-los em um local compartilhado para que as pessoas se lembrem de como acessá-los.

Para APIs e materiais de entrega de machine learning, a consistência está no projeto do produto. À medida que o portfólio de APIs e modelos da equipe de ciência de dados cresce, pode ser extremamente difícil acompanhar como cada um funciona. Quanto mais consistentes forem as APIs, mais facilmente poderão ser usadas. As regras de consistência incluem:

- *Consistência na entrada* – a forma como seus modelos e APIs recebem dados deve seguir o mesmo formato tanto quanto possível. Todos podem receber objetos JSON com os mesmos nomes de parâmetro, por exemplo.
- *Consistência na saída* – a forma como a saída é estruturada deve fazer sentido com a forma como a entrada é estruturada, assim como o restante das APIs criadas pela equipe funciona. Se o modelo recebe JSON como entrada, ele deve retornar JSON como resultado.
- *Consistência na autenticação* – é provável que os modelos e APIs exijam alguma forma de autenticação para segurança. Qualquer que seja o método utilizado, ele deve ser consistente no maior número de APIs, especialmente porque é fácil perder o controle de quais são as credenciais para cada API.

Além de ajudar os stakeholders, toda essa consistência será útil para você! Quanto mais puder padronizar todas essas partes do trabalho de ciência de dados, menos terá de pensar sobre elas (e mais poderá se concentrar nas partes interessantes). Quanto mais uma equipe de ciência de dados puder padronizar, mais fácil será transmitir o trabalho a pessoas diferentes. A padronização é positiva para todos.

**Elizabeth Hunter, vice-presidente sênior de implementação de  
estratégia tecnológica da T-Mobile: como gerenciar  
relacionamentos**

O relacionamento interpessoal é importante, mas, por vezes, negligenciado de toda interação corporativa. As pessoas procuram de forma subconsciente indicações sociais e emocionais para gerar estabilidade, conforto e conexão; quando encontram essas indicações tendem a ser mais abertas às ideias e a experiências novas. Essa conexão com uma pessoa proporciona um ambiente acolhedor para toda informação que deve lhes apresentar. Entretanto, estabelecer um bom relacionamento com alguém não é tão claro como outras tarefas de trabalho. Algumas pessoas formam rapidamente conexões de forma amigável, enquanto outras levam muito mais tempo e requerem inúmeras interações com profunda discussão pessoal.

Ao longo dos anos, descobri que grande parte do sucesso na minha carreira dependia dos relacionamentos que dediquei tempo para estabelecer, quer se tratasse do apoio de alguém em uma reunião importante, um executivo ouvindo uma das minhas ideias ou de alguém me oferecendo uma nova oportunidade. Grande parte do meu crescimento na carreira foi trabalhar arduamente e demonstrar o valor de meu trabalho, mas ter desenvolvido conexões com as pessoas mostrou-lhes um pouco sobre mim, o que poderiam esperar, o quanto de margem de manobra estavam dispostos a me dar e quanto confiariam em mim sem precisar de provas.

Sendo uma pessoa introvertida, desenvolver relações é algo menos natural para mim do que para outras, e tive de trabalhar isso. Retrospectivamente, percebo agora que realizei uma série de pequenos experimentos – formei hipóteses com base no que sabia ou observei como eu poderia me conectar com alguém, testei se funcionou ou não, fiz ajustes com base em novas informações que aprendi sobre uma pessoa, repeti em minhas interações até descobrir do que gostavam e consegui me conectar bem com elas. Não quer dizer que você deve mudar em função dos outros, mas alguma empatia com o que torna os outros confortáveis ajuda bastante.

## 12.3 Como priorizar o trabalho

Como cientista de dados que tenta apoiar uma empresa, muitas vezes é preciso decidir em qual tarefa trabalhar. Embora algumas equipes tenham um gerente de projeto que decida sobre o trabalho de cada cientista de dados, ainda deve recomendar qual seria a tarefa seguinte. Essas tarefas podem variar muito quanto ao tópico e ao escopo, e cada uma pode vir de um stakeholder diferente. É possível classificar as tarefas em três categorias:

- *Tarefas rápidas que vêm diretamente dos stakeholders* – essas tarefas tendem a ser pequenas solicitações, como: “Faça um gráfico de vendas do período”. Muitas vezes, são urgentes e, como não demoram muito tempo, é difícil negar. Porém, é uma distração do trabalho mais importante e, à medida que os pedidos se acumulam, fica cada vez mais difícil ser produtivo.
- *Projetos de longo prazo para a empresa* – esses projetos são a parte central do trabalho de um cientista de dados. A construção de dashboards, a realização de análises de longo prazo e a criação de modelos para colocar em produção recaem nesta categoria. Essas tarefas tendem a ser muito importantes, mas como podem levar semanas ou meses para serem concluídas, pois não são sempre urgentes.
- *Ideias que você acha terem um benefício a longo prazo* – dada a natureza da ciência de dados, elas geralmente são mais técnicas, como a

criação de um modelo de machine learning para prever quando um cliente chamará o suporte antes que o faça. Essa categoria também inclui o trabalho que o torna mais produtivo, como criar funções ou até mesmo uma biblioteca para resolver problemas comuns mais rapidamente. Se um processo manual levar horas por semana para ser executado, poderá automatizar a tarefa, não oferecendo benefício direto para a empresa, mas se liberando indiretamente para fazer mais. Ninguém está pedindo esse trabalho, mas parece importante.

Pode ser difícil descobrir a tarefa mais importante para trabalhar e o que deixar para depois, especialmente quando várias pessoas estão lhe solicitando trabalhos. Ao mesmo tempo, o trabalho que é muito importante para os stakeholders pode não ser relevante para a empresa como um todo. Como cientista de dados, raramente você tem a capacidade de recusar solicitações dos stakeholders, porque geralmente são as pessoas que comandam a empresa. Tudo isso cria um ambiente no qual a decisão sobre o que trabalhar pode influenciar enormemente a empresa; também restringe o que você pode escolher fazer.

Esta é uma área com a qual muitos cientistas de dados têm dificuldade. Quando os stakeholders fazem solicitações, é natural querer agradá-los e não os decepcionar. Ao lado disso, as solicitações que fazem podem ser intelectualmente interessantes. No entanto, tentar atender a todas as solicitações é insustentável, pois são intermináveis. Ademais, responder a uma questão com dados muitas vezes conduz a novas questões, razão pela qual atender às solicitações cria ainda mais trabalho em vez de reduzir a sua quantidade.

Quando estiver considerando possíveis tarefas para trabalhar, algo que ajuda é se concentrar em duas perguntas:

- *Este trabalho terá algum impacto?* Saber o resultado dessa análise afeta materialmente a empresa? Algum resultado mudaria as decisões? Este modelo de machine learning aumentaria o lucro?
- *Este trabalho fará algo novo?* Ele seria aplicado a um processo existente repetidamente ou seria tentar algo diferente?

As respostas a essas duas questões criam quatro combinações de tipos de

trabalho:

- Inovador e com impacto.
- Não inovador, mas com impacto.
- Inovador, mas sem impacto.
- Nem inovador nem com impacto.

Nas próximas seções, serão detalhadas cada uma dessas combinações.

### **12.3.1 Trabalho inovador e com impacto**

O trabalho inovador que altera a empresa é o que a maioria dos cientistas de dados quer passar fazendo em sua carreira. Um projeto de exemplo seria algo como pegar dados de estoque que nunca foram tocados por um cientista de dados e usar um modelo de machine learning moderno para otimizar o produto de pedidos, economizando milhões de dólares da empresa. São do tipo de projetos que, no melhor cenário, são publicados em revistas como *Harvard Business Review* ou *Wired*.

Infelizmente, poucos projetos de empresas recaem nessa categoria. Para existir, esses projetos precisam de muitas coisas:

- É necessário que haja dados suficientes para que os métodos de ciência de dados sejam úteis.
- Tem de haver um indício interessante nos dados para que modelos possam aprender.
- O setor da empresa tem que ser grande ou importante o suficiente para que as mudanças possam fazer a diferença (e não simplesmente otimizar o estoque de canetas de quadro branco para o escritório).
- O problema deve ser complexo ou único o suficiente para que as pessoas não o tenham tentado antes.

O conjunto de problemas em uma empresa que recai em todas essas categorias é extremamente pequeno.

Esses projetos são grandes porque geram entusiasmo entre os stakeholders e os cientistas de dados. Os stakeholders sentem-se ótimos porque podem ver o evidente valor do projeto. Os cientistas de dados têm vontade de tentar novos métodos em novos dados e ver os resultados. Se encontrar um

projeto nessa categoria, faça tudo o que puder para cuidar dele. Esses projetos são do tipo que pode definir uma carreira, mas, como têm tantas exigências, são muito raros e, muitas vezes, não são bem-sucedidos.

### **12.3.2 Trabalho não inovador, mas com impacto**

Esses projetos não são inovadores, mas alteram a empresa, como uma análise de dados muito simples que convence uma equipe a lançar determinado produto. Muitas vezes, essa persuasão equivale a fornecer provas de que todos os indícios são verdadeiros; não é particularmente inovador, mas ajudará a empresa. Na engenharia, esses projetos poderiam estar tomando um modelo que já foi implantado em uma divisão da empresa e o reimplantando em outra divisão. Outro tipo de trabalho que se enquadra nessa categoria é simplificar tarefas que levam muito tempo. Esse tipo de trabalho não é inovador, mas melhora a empresa.

Embora o trabalho não seja glamuroso, o importante é que ele ajude a empresa. Ajudar os stakeholders a ver o valor do trabalho de ciência de dados é excelente para obter aprovação. Se tiver mais aprovação, na próxima vez que um projeto superar o orçamento ou não funcionar como o esperado, é mais provável que as pessoas continuem confiando em você. Tanto quanto possível, tente assumir esses projetos.

No fim do dia, o trabalho de um cientista de dados é fornecer valor à empresa, e não fazer o trabalho mais fascinante. Uma competência valiosa para os cientistas de dados é poder tolerar esse tipo de trabalho de importância, mas não interessante. Dito isso, se um emprego estiver completamente cheio de projetos que não lhe interessam nem o ensinam, é completamente apropriado procurar um emprego diferente. É totalmente válido considerar a satisfação no trabalho quando você está priorizando o trabalho; só garanta de que não seja a única consideração.

### **12.3.3 Trabalho inovador, mas sem impacto**

Esse trabalho é inovador, mas não é útil para a empresa, como a pesquisa de novos algoritmos e métodos teóricos de ciência de dados que têm poucas chances de serem usados. Tais projetos podem ser torres de marfim, nas quais as pessoas passam meses ou anos trabalhando sem interagir com

outros grupos e que acabarão não sendo utilizados. Podem demandar muito tempo das equipes de ciência de dados e acabam custando milhões de dólares com pouco retorno. Apesar disso, esses projetos atraem cientistas de dados, como mariposas na luz.

Esses projetos tendem a começar dentro da equipe de ciência dos dados e se concentram no que é metodologicamente interessante, em vez de no que é útil para a empresa. Um cientista de dados pode ter lido um artigo científico descrevendo uma nova técnica teórica e convencer os demais integrantes da equipe que eles *devem* tentar usá-la em seus próprios dados. Seis meses depois, é claro que o método não é tão bom quanto o artigo sugeriu, e, mesmo que fosse, ninguém na empresa necessita particularmente dos resultados que o algoritmo teria fornecido. Pior ainda, o cientista de dados está lendo um novo artigo, e o processo se repete...

Muitas vezes, os stakeholders sequer sabem sobre esses projetos. No máximo, notam que alguns dos cientistas de dados da equipe parecem estar muito ocupados trabalhando em algo que soa muito difícil, mas ninguém explicou o que é. Como cientista de dados, é fácil pensar que ao completar um projeto, as pessoas poderão encontrar um uso para ele. Na prática, se não puder ver de imediato uma utilização para o projeto, os stakeholders provavelmente também não encontrarão. Na medida do possível, não fique preso a tais projetos, os quais podem não contribuir para a empresa, fazendo com que as pessoas questionem seu valor.

### **12.3.4 Trabalho nem inovador nem impactante**

Infelizmente, muitas das solicitações de trabalho recebidas pelos cientistas de dados não são inovadoras nem têm algum impacto. O exemplo mais clássico é um relatório frequentemente atualizado que não é automatizado e demora muito tempo para ser feito, mas ninguém o lê sempre que é publicado. Esse tipo de trabalho demanda muito tempo e esforço, mas, se for entregue a muitos stakeholders, ninguém está disposto a ser aquele stakeholder que toma coragem para dizer que o trabalho já não precisa mais ser feito. À medida que a empresa coleta mais e mais relatórios necessários ao longo do tempo, o tempo exigido para gerá-los pode eventualmente pesar em uma equipe de ciência de dados.



Embora o relatório seja um tipo de trabalho que tem o potencial de não ser inovador nem impactante, muitas pequenas solicitações únicas podem ser incluídas nessa categoria. Os executivos que gostam de dados e gráficos podem reiteradamente fazer pedidos para a equipe de ciência de dados, como: “Faça um gráfico das vendas semanais na Europa” ou “Encontre o produto que tem a maior queda nos pedidos nas últimas 12 semanas”. Nenhuma dessas solicitações pode ser particularmente difícil, mas, juntas, demandam tempo e provavelmente não fornecem muito valor para a empresa.

Esses tipos de situações são difíceis porque as respostas não são fáceis. Você pode tentar automatizar relatórios e processos que levam muito tempo, mas essa tarefa em si leva também muito tempo, e você pode obter apenas melhorias limitadas, dependendo da tecnologia que está sendo usada. Se os stakeholders de alto nível estão fazendo repetidas solicitações, por exemplo, é difícil negar sem comprometer a posição da equipe de ciência de dados.

Apesar dessas situações difíceis, é sua responsabilidade, enquanto cientista de dados, tentar defender que seu tempo seja bem utilizado. Se muitas dessas tarefas estiverem acontecendo, deve deixar claro a outras pessoas que essas tarefas podem não valer a pena, conversando com seu gerente ou com os stakeholders. É provável que já saibam que esse trabalho não é especialmente útil, mas tendo conversas contínuas sobre os processos e como acham que devem ser melhorados, as pessoas estarão menos dispostas a aceitar o status quo. Se não, às vezes o melhor a se fazer é tentar fazer suas melhorias sugeridas primeiramente e depois mostrá-las.

## **12.4 Observações finais**

Trabalhar com stakeholders é um processo constante ao longo de um projeto. Você precisa entender as suas necessidades e por que estão fazendo as solicitações. Um projeto começa devido a uma solicitação do stakeholder, mas o que ele está solicitando provavelmente mudará durante o projeto em si, e é sua responsabilidade acompanhar as mudanças. Quanto mais puder alinhar seu projeto com o que o stakeholder está solicitando, é menos provável que o projeto falhe. No Capítulo 13, falaremos sobre o que

acontece quando um projeto de ciência de dados falha, como quando a comunicação com os stakeholders é interrompida.

## **Sam Barrows, cientista de dados da Airbnb: como transformar solicitações em diálogos**

Uma ferramenta valiosa para trabalhar com stakeholders é transformar as solicitações em diálogos. Muitas vezes, os colegas podem solicitar que você conclua tarefas específicas. Em vez de aceitar ou rejeitar imediatamente essas solicitações, inicie um diálogo sobre o motivo pelo qual a solicitação está sendo feita. Qual necessidade da empresa essa solicitação procura resolver? Existe uma forma melhor de alcançar o resultado pretendido? Ao compreender as motivações por trás das solicitações recebidas, é mais provável que faça um trabalho significativo.

Essa estratégia está em praticar a negociação com base em interesses, onde os stakeholders, em uma negociação, concentram-se em abordar os interesses subjacentes, em vez de apenas as suas necessidades mais imediatas. Nesse caso, as necessidades imediatas são as solicitações recebidas, enquanto os interesses subjacentes são as motivações de negócios por trás dessas solicitações, bem como os objetivos da equipe de ciência de dados.

## **12.5 Entrevista com Sade Snowden-Akintunde, cientista de dados da Etsy**

Sade Snowden-Akintunde trabalha na Etsy, onde se especializa em design e análise de experimentos para melhorar as experiências de compra de consumidores internacionais. Suas áreas de especialização incluem experimentos e testes A/B, implementação de práticas de dados confiáveis e dimensionamento da infraestrutura de dados.

### **Por que é importante gerenciar os stakeholders?**

Infelizmente, não importa quão inteligente você seja se não conseguir comunicar conceitos a stakeholders não técnicos. No fim do dia, muitas dessas empresas são administradas por pessoas que podem não ter o mesmo nível de competência técnica que você. É preciso poder se comunicar com eles de uma maneira que os faça sentirem-se capazes e que permita se defenderem, caso necessário. Gerenciar os stakeholders é, provavelmente, um dos aspectos mais importantes da ciência de dados, mas, muitas vezes, é

o menos destacado.

## **Como aprendeu a gerenciar stakeholders?**

Por tentativa e erro: tive situações que funcionaram e situações que não funcionaram, e prestei atenção. O maior ensinamento que tive foi perceber a importância de comunicar-se logo e repetir as coisas para garantir que as pessoas compreendam o que está dizendo. No início da minha carreira na área da ciência de dados, presumi que, se falasse algo uma vez e alguém concordasse, então sabiam exatamente do que eu estava falando, mas as pessoas podem nem saber que não entendem o que você está falando.

## **Houve algum momento em que teve dificuldade com um stakeholder?**

No início da minha carreira, tinha receio em me defender e defender minha perspectiva como designer de experimento. Outras pessoas faziam experimentos sem sentido para mim, mas não dizia nada. Depois eu tentava analisar os experimentos concluídos, e os resultados seriam difíceis de interpretar por causa do design do experimento. Eu deveria ter trabalhado com os stakeholders desde o início para comunicar como eu poderia analisar melhor o trabalho deles e o que eles iriam obter a partir do experimento deles com um design adequado. Percebi que tenho que dizer algo bem no início se quiser ser capaz de realizar meu melhor trabalho no final.

## **Com o que os cientistas de dados iniciantes frequentemente se enganam?**

Penso que os cientistas de dados iniciantes presumem que as pessoas vão reconhecer automaticamente o valor do trabalho deles. Isso é especialmente comum entre cientistas de dados com bagagem acadêmica. Tendemos a seguir tudo à risca, inclusive o método científico. Embora seja importante no meio acadêmico, apenas trabalhar muito não é necessariamente o que fará com que as pessoas reconheçam o valor do seu trabalho. A maneira como se comunica é o que faz as pessoas reconhecerem o valor de seu trabalho.

## **Você tenta sempre explicar a parte técnica da ciência de dados?**

Depende de quanto o stakeholder quer ser envolvido. Trabalhei com gerentes de projeto que não queriam se envolver em nada técnico. Se eu dissesse apenas: “Isso não está funcionando agora”, já era suficiente. Também trabalhei com gerentes de projetos que queriam saber todos os detalhes e o que descobri é que eles tendem a ficar um pouco perdidos. Algumas pessoas querem confirmações regulares e que sejam informados sobre o que está acontecendo, e, mesmo que estejam conscientes de que não entendem, eles só querem se sentir informados. Por isso, garanto que se sintam informados.

## **Qual é seu último conselho para aspirantes ou cientistas de dados iniciantes?**

Penso que as pessoas tendem a querer entrar em carreiras técnicas como a ciência de dados, por pensarem que podem se concentrar no elemento lógico e não lidar com o elemento humano. Mas não é o caso. Quando as pessoas estão considerando uma carreira na ciência de dados, elas devem realmente pensar se estão dispostas a terem um ego pequeno, a fim de se comunicarem e fazer bem o seu trabalho. É muito fácil dizer: “Quero aprender a construir esse modelo, quero aprender a testar A/B e também todas essas coisas técnicas”. Embora seja excelente, as competências sociais são o que o levarão muito longe na carreira.

## **Resumo**

- Os stakeholders surgem sob muitas formas, com muitas necessidades.
- Crie relacionamentos com os stakeholders para que possam confiar de maneira consistente em você.
- Tenha uma comunicação contínua e mantenha os stakeholders informados sobre cronogramas e dificuldades com os projetos.

## **Recursos dos capítulos 9–12**

**Livros**

*Beautiful Evidence*, de Eduard Tufte (Graphics Press)

Eduard Tufte é uma lenda no campo da visualização de dados, e seus livros são repletos de orientações detalhadas sobre como pensar gráficos e tabelas. Ele também tem outras publicações; você pode adquirir essa bibliografia em conjunto ou, ainda melhor, fazer um dos cursos de um dia que ele faz como um passeio. Uma palavra de cautela, porém: o conselho dele às vezes é acadêmico. É praticamente impossível fazer tudo o que ele sugere e ter tempo para realizar qualquer outra parte de seu trabalho, além de fazer visualizações.

*Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*, de Claus O. Wilke (O'Reilly Media)

Se Eduard Tufte oferece uma visão acadêmica de visualizações, Wilke fornece a versão prática e aplicada. O texto explica como pensar sobre decisões de visualização no dia a dia. Quando gráficos de caixa (boxplots) são bons? Os gráficos tipo pizza são tão ruins quanto as pessoas dizem? Este livro irá guiá-lo nessas decisões.

*The Pyramid Principle: Logic in Writing and Thinking*, de Barbara Minto  
(Trans-Atlantic Publications)



Este livro está há anos sem reedição (embora possa encontrar cópias usadas), mas ainda é recomendado como a obra-base sobre comunicar-se bem. Minto explica como pensar na estruturação de um relatório ou de uma apresentação para que o público o entenda, orientando criticamente, tal como ordenar o conteúdo de uma forma que seja significativa, não apenas na ordem de criação. Minto é ex-consultora da prestigiada empresa de consultoria McKinsey, e o livro está repleto de lições que os consultores dominam.

*The Design of Web APIs*, de Arnaud Lauret (Manning)

É comum aprender a projetar APIs com a experiência; no fim, faz o suficiente delas para que os projetos comecem a ser sensatos. Este livro é um atalho desse processo. Começa por definir o que são APIs e como são estruturadas; em seguida, passa pelo design delas e pelas melhores práticas. Cobre até tópicos como a documentação do OpenAPI para que possa escrever especificações compartilháveis para suas APIs.

*Amazon Web Services in Action*, 2ª ed, de Michael Wittig e Andreas Wittig (Manning)

*Azure in Action*, de Chris Hay e Brian H. Prince (Manning)

*Google Cloud Platform in Action*, de JJ Geewax (Manning)

Esses três livros tratam de como utilizar o Amazon Web Services, Microsoft Azure e Google Cloud Platform, respectivamente. À medida que aprende a implantar modelos de machine learning, é bom ter um lugar para hospedá-los, sendo esses três provedores de nuvem as principais opções. É possível escolher qualquer plataforma que soar mais útil para você e ler o livro apropriado para aprender o básico.

*Difficult Conversations: How to Discuss What Matters Most*, de Douglas Stone et al. (Penguin Publishing)

A comunicação é sempre delicada, mas é ainda mais difícil quando o tópico é complicado ou as pessoas têm muito interesse. Esse livro é sobre tratar de assuntos que as pessoas via de regra evitam. O livro fornece um grande conjunto de competências para um cientista de dados porque, muitas vezes, eles têm de apresentar resultados insatisfatórios às pessoas com as quais trabalham.

*Getting to Yes: Negotiating Agreement Without Giving In*, de Roger Fisher, William L. Ury e Bruce Patton (Penguin Publishing)

Ser um cientista de dados requer muita negociação, desde persuadir uma equipe a conceder acesso a dados até instar um executivo a prestar atenção nos seus achados. Ser capaz de convencer e negociar com sucesso nesses momentos pode ser mais importante para seu sucesso do que qualquer competência técnica. *Getting to Yes* é um grande recurso para aprender a como negociar com os stakeholders e a obter os resultados que deseja.

*Software Requirements*, 3a. ed, de Karl Wieggers e Joy Beatty (Microsoft Press)

Definir o que é necessário para um projeto de uma forma que possa ser entendida pelo negócio é difícil. Este bem recomendado livro aborda sobre como criar requisitos e geri-los ao longo do projeto. Embora juntar requisitos não seja a parte mais glamurosa da ciência de dados, pode fazer ou quebrar a execução de um projeto.

## **Blogs**

“R in Production”, de Jacqueline Nolis e Heather Nolis

<http://mng.bz/YrAA>

Essa série de três partes aborda a criação de uma API em R com o pacote Plumber, sua implementação como um contêiner Docker e a preparação para a empresa. O contêiner R Docker de código aberto fornecido está em uso pela T-Mobile.

“Advice for new and junior data scientists: what I would have told myself a few years ago”, de Robert Chang

<http://mng.bz/zlyX>

Nesse post popular, Robert Chang, nosso entrevistado do Capítulo 1, fala dos seis princípios fundamentais que aprendeu na sua trajetória para se tornar um cientista de dados sênior na Airbnb. Essas informações importantes podem demorar anos para serem aprendidas sozinho, por isso aproveite o atalho e comece a aplicá-las agora.

“Data science foundations: know your data. Really, really, know it”, de Randy Au



<http://mng.bz/07Pl>

Randy Au, nosso entrevistado do Capítulo 2, dá o seguinte conselho a cada pessoa nova nos dados: “Conheça seus dados, de onde vêm, o que está neles, o que significam. Tudo começa aí”. Nesse post, ele define como conhecer seus dados, desde o início do layout até saber as decisões de coleta sendo tomadas.

“How to work with stakeholders as a data scientist: what I would have told myself when I started”, de Sam Barrows

<http://mng.bz/KEPZ>

Compartilhamos a primeira das sete sugestões de Sam para trabalhar produtivamente com stakeholders como uma barra lateral no Capítulo 12 (“Como transformar solicitações em diálogos”), mas as outras seis merecem ser lidas.

## PARTE IV

# Como crescer em sua função em ciência de dados

O conteúdo desta Parte IV do livro é para ser usado quando você já estiver confortavelmente ocupando um cargo da ciência de dados, pois trata sobre o que vem em seguida. Os tópicos aqui tratados afetam em algum momento todos os cientistas de dados, embora os temas não sejam discutidos com frequência. É fácil supor que, se tiver um trabalho estável de ciência de dados, você chegou ao topo da carreira, mas sempre há mais para aprender. O objetivo desta parte final é fornecer materiais para ajudá-lo a passar de um cientista de dados júnior para um cientista de dados sênior e ainda mais além.

O Capítulo 13 trata sobre como lidar com falhas em projetos de ciência de dados. Esse tópico é extremamente importante para os cientistas de dados experientes, porque, à medida que a carreira progride, com certeza haverá falhas. O Capítulo 14 mostra sobre como juntar-se à comunidade da ciência de dados, desde a redação de posts de blog até a participação em conferências. Embora não seja necessário para os cientistas de dados, o envolvimento da comunidade pode ser extremamente benéfico para a criação de uma rede e para conseguir futuros empregos. O Capítulo 15 aborda a difícil tarefa de sair de um cargo de ciência de dados da melhor forma para sua carreira. Como último capítulo, o 16 discute alguns dos principais caminhos de carreira após tornar-se cientista de dados sênior, bem como tornar-se um gerente ou uma liderança técnica.

## CAPÍTULO 13

# Quando seu projeto de ciência de dados falha

Este capítulo abrange:

- Por que os projetos de ciência de dados tendem a falhar
- O que fazer quando seu projeto falhar
- Como lidar com as emoções negativas em caso de falhas

A maioria dos projetos de ciência de dados trata de empreendimentos de alto risco. É uma tentativa de presumir algo que ninguém previu antes, otimizar algo que ninguém otimizou antes ou entender dados que ninguém examinou antes. Não importa o que estiver fazendo, você é a primeira pessoa a fazê-lo; o trabalho é quase sempre exploratório. Como os cientistas de dados estão continuamente fazendo coisas novas, chegará inevitavelmente um momento em que irão se deparar com a situação de que o esperado simplesmente não é possível. Todos temos de lidar com nossas ideias que não tiveram êxito. O fracasso é doloroso e angustiante; você quer parar de pensar sobre a ciência de dados e fica devaneando sobre como abandonar radicalmente a área.

Como exemplo, considere uma empresa que constrói um modelo de machine learning para recomendar produtos em seu site. Provavelmente os eventos começam com algumas reuniões nas quais a equipe de ciência de dados convence os executivos de que o projeto é uma boa ideia. A equipe acredita que, ao utilizar informações sobre os clientes e as respectivas transações, em seguida pode prever o que os clientes pretendem comprar. Os executivos compram a ideia e o projeto recebe sinal verde. Muitas outras empresas têm esses modelos, que parecem simples; portanto, o projeto deve funcionar.

Infelizmente, quando a equipe começa a trabalhar, a realidade é outra. Talvez descubram que, como a empresa recentemente substituiu os sistemas, os dados de transações estão disponíveis somente para os últimos meses. Ou talvez a equipe conduza um experimento e descubra que as pessoas que veem o motor de recomendação não compram nada a mais do que as aquelas que não o veem. Problemas como esses se acumulam; no fim, a equipe abandona o projeto, desanimada.

Neste capítulo, definimos um projeto como *falho* quando não cumpre o seu objetivo. No caso de uma análise, o projeto pode falhar quando não auxilia o stakeholder a responder à questão da empresa. Para um problema de machine learning na produção, um projeto pode falhar quando não for implantado ou não funcionar quando implantado. Os projetos podem falhar de muitas maneiras.

Os cientistas de dados tendem a não falar de falha nos projetos, embora isso ocorra com frequência. Quando um projeto falha, o cientista de dados pode se sentir vulnerável. Se seu projeto falhar, poderá pensar que: “Se fosse um cientista de dados melhor, isso não teria acontecido”. Poucas pessoas sentem-se confortáveis compartilhando histórias sobre questionar suas próprias competências.

Em sua essência, ciência de dados é pesquisa e desenvolvimento. Diariamente, cientistas de dados pegam dados que nunca foram analisados e procuram uma tendência que possa estar presente ou não. Os cientistas de dados constroem modelos de machine learning com dados que podem não ter um sinal. É impossível que essas tarefas sejam sempre bem-sucedidas, pois novas tendências e sinais são raramente encontrados em qualquer área. No entanto, em uma área como a engenharia de software, em geral é possível concluir uma tarefa (embora possa levar mais tempo e recursos do que o planejado).

É importante compreender como os projetos de ciência de dados falham e o que fazer quando isso acontece. Quanto melhor entender um projeto que falhou, mais falhas futuras poderão ser evitadas. Os projetos com falhas também podem fornecer informações sobre o que será bem-sucedido, investigando quais partes do projeto funcionaram. Com um pouco de trabalho, é possível transformar um projeto com falhas em algo que poderia

ser útil na empresa.

Neste capítulo foram abordados três tópicos: por que os projetos de ciência de dados falham, como pensar o risco do projeto e o que fazer quando um projeto apresenta falhas. Discutimos as três razões principais pelas quais a maioria dos projetos falha, o que fazer com o projeto e como lidar com as emoções que pode sentir se o projeto falhar.

## **13.1 Por que seu projeto de ciência de dados falha?**

Parece que os projetos de ciência de dados falham devido a uma lista interminável de razões. Do orçamento à tecnologia e até tarefas que levam muito mais tempo do que o esperado para serem concluídas, há muitas razões para as falhas. Esses muitos tipos de falhas dividem-se em alguns temas centrais.

### **13.1.1 Os dados não são aqueles que você queria**

Não é possível analisar todas as fontes de dados possíveis antes de iniciar um projeto. É muito importante fazer suposições fundamentadas sobre o que está disponível com base no que conhece da empresa. Quando o projeto começa, muitas vezes você descobre que muitas de suas suposições não são verdadeiras. Talvez os dados não existam, não estejam armazenados em um formato útil ou não estejam armazenados em um local de fácil acesso. Se estiver fazendo uma análise para entender como a idade de um cliente afeta o uso de um programa de fidelidade, por exemplo, talvez descubra que os clientes nunca são questionados sobre suas idades ao se juntarem ao programa. Essa falha pode acabar imediatamente com um projeto.

**Exemplo de falha: análise do status do programa de fidelidade**

Um diretor do departamento de marketing de uma grande rede de restaurantes quer entender se os clientes gastam de forma diferente à medida que aumentam o status no programa de fidelidade da empresa. O programa tem os níveis prata, ouro e platina, e o diretor quer saber se quem atinge o nível platina comprou da mesma forma quando estava ainda no nível prata.

A equipe de ciência de dados concorda em analisar essa solicitação porque a tarefa parece bastante simples, e eles não trabalharam antes com os dados de fidelidade. Eles ficam chocados ao descobrir que o antigo banco de dados de programas de fidelidade não acompanha os níveis históricos de programas – exatamente onde os clientes estão agora. Se um cliente estiver agora no nível platina, não há como saber quando estavam no nível prata ou ouro. Portanto, é impossível fazer a análise.

A equipe de ciência de dados recomenda que o sistema seja ajustado, mas alterar uma arquitetura de banco de dados de programas de fidelidade requer milhões de dólares, e há pouca demanda para isso na empresa; portanto, nenhuma mudança é feita e a ideia de análise é abandonada.

Como você precisa de dados antes de fazer qualquer coisa, esses são os primeiros grandes problemas que surgem. Uma reação comum ao se deparar com um problema desses é a negociação interna, na qual você tenta trabalhar com as lacunas nos dados, e diz coisas do tipo: “Bem, não temos dados de uma década como queríamos, mas talvez um ano de dados será suficiente para o modelo” e, então, torce pelo melhor. Às vezes, essa abordagem pode funcionar, mas as soluções alternativas nem sempre são adequadas para a viabilização do projeto.

Quando um projeto começa, nem sempre se tem acesso aos dados ou até mesmo nem uma compreensão completa do que ele se trata (um problema especial na consultoria, onde não se tem acesso aos dados até que o trabalho para o projeto seja vendido). Além disso, os dados podem existir, mas têm uma falha crítica que os torna inúteis. Os dados podem existir em uma tabela de banco de dados, mas as identidades de clientes podem ficar corrompidas e inutilizáveis. Há tantas maneiras de um conjunto de dados ter problemas que é extremamente difícil verificá-los em sua totalidade antes de iniciar um projeto. Por essa razão, é comum que os projetos de ciência de dados mal passem da fase de lançamento.

Quanto mais rápido conseguir acesso aos dados e explorá-los, mais rápido poderá reduzir o risco de dados inadequados. O melhor cenário para evitar esse erro é obter amostras de dados antes de iniciar um projeto. Se não for viável, o próximo melhor cenário é ter um cronograma de projeto concebido em torno da possibilidade de que os dados não sejam adequados. Tendo dado um primeiro passo no projeto para aceitá-lo ou não, no qual os stakeholders concordam em reavaliar a viabilidade do projeto, há menos chances de os stakeholders se surpreenderem com possíveis dados inadequados.

Caso enfrente dificuldade com a falta de dados bons, suas opções são limitadas. Também pode tentar, por exemplo, encontrar fontes de dados



alternativas para substituí-los. Talvez não tenha dados sobre quais produtos foram comprados, mas sabe o volume de produto fabricado e pode usá-lo. O problema geralmente é que esses dados alternativos são diferentes o suficiente para causar problemas reais com a análise.

Quando não consegue encontrar um substituto viável, tudo o que se pode fazer, às vezes, é iniciar um projeto paralelo para começar a coletar dados melhores. Adicionar instrumentação e telemetria a sites e aplicativos, criar bancos de dados para armazenar dados em vez de jogá-los fora e realizar outras tarefas pode ajudar a equipe a assumir a tarefa no futuro com melhores dados coletados.

### **13.1.2 Os dados não têm um sinal**

Suponha que um apostador contrate um cientista de dados esperando usar estatísticas para ganhar um jogo de dados. O jogador joga um dado de seis lados 10 mil vezes e registra as jogadas; depois, contrata o cientista de dados para criar um modelo que irá prever o próximo resultado ao jogar o dado. Apesar de o cientista de dados ter uma enorme quantidade de dados, não há como prever qual será o resultado, além de atribuir a cada lado uma probabilidade de  $1/6$  (se o dado for justo). Apesar de o cientista de dados ter muitos dados (informações), não há nenhum sinal neles a respeito de qual será o próximo resultado.

Esse problema de não ter um indício nos dados é extremamente comum na ciência de dados. Suponha que esteja executando um site de comércio eletrônico e queira criar um modelo para prever quais clientes encomendarão produtos com base no navegador, dispositivo e sistema operacional. Não há como saber antes de iniciar um projeto se esses pontos de dados poderiam realmente ser usados para prever se o cliente fará um pedido ou se os dados não têm um indício, da mesma forma que em um jogo de dados, em que saberemos o resultado somente após jogar o dado. O ato de criar um modelo de machine learning para fazer uma predição é testar os dados para ver se há algum sinal neles, que pode muito bem não existir. Na verdade, em muitas situações, seria mais surpreendente que *houvesse* um sinal do que *não* haver um sinal.

## **Exemplo de falha: como detectar bugs em um site com dados de vendas**

Uma hipotética empresa de comércio eletrônico tem um problema: o site continua apresentando erros e bugs. Pior ainda: os erros nem sempre são detectados pelo DevOps ou pela equipe de engenharia de software. Uma vez, o erro foi detectado pela equipe de marketing, a qual observou que a receita diária estava muito baixa. Quando o marketing detecta um bug em vez do DevOps ou da engenharia, trata-se de uma situação desfavorável.

A equipe de ciência de dados propõe-se a usar técnicas estatísticas de controle de qualidade nos dados de vendas para que possa alertar quando a receita for tão baixa que deve existir um bug no site. Eles têm uma lista de dias em que foram detectados os bugs e os dados históricos de receita. Parece simples usar as vendas para prever bugs.

Infelizmente, as razões pelas quais a receita pode mudar diariamente torna a detecção de bugs quase impossível. A receita pode ser baixa por causa do dia da semana, do momento do ano, das promoções do marketing, dos eventos globais ou de qualquer outra coisa. Embora o marketing já tenha sido capaz de detectar um bug, esse fato não era generalizável porque não havia um sinal para ele nos dados.

Não ter um indício nos dados pode, infelizmente, ser o fim do projeto. Se um projeto for feito na tentativa de encontrar um relacionamento nos dados e fazer uma predição com base neles, e não existir um relacionamento, a predição pode não ser feita. Uma análise pode não trazer algo novo ou interessante ou um modelo de machine learning pode não ter quaisquer resultados que sejam melhores do que por chance aleatória.

Se não conseguir encontrar o sinal nisso tudo, há algumas alternativas:

- *Reformular o problema.* É possível tentar reformular o problema para ver se há um indício diferente. Suponha que tenha um conjunto de artigos e esteja tentando prever o artigo mais relevante para o usuário. Seria possível formular como um problema de classificação para tentar classificar qual artigo em um conjunto de artigos é o mais relevante.
- *Alterar a fonte de dados.* Se nada parecer mostrar um indício nos dados, é possível tentar alterar a fonte desses dados. Assim como a falha anterior aponta não ter bons dados, adicionar uma nova fonte de dados ao problema às vezes cria um indício inesperado. Infelizmente, em geral se começa com o conjunto de dados que tinha a maior probabilidade de ser útil; portanto, a probabilidade de que essa estratégia venha a salvá-lo é relativamente limitada.

É normal que os cientistas de dados que estão presos nessa situação tentem utilizar um modelo mais poderoso para encontrar um sinal. Se uma regressão logística não puder fazer uma predição significativa, eles tentam um modelo de floresta aleatória (random forest). Se um modelo de floresta aleatória não funcionar, eles tentam uma rede neural. Cada método acaba sendo mais demorado e mais complexo do que o outro. Embora esses métodos possam ser úteis para obter predições mais precisas, não é possível criar algo do nada.

Na maioria das vezes, se o método mais simples não conseguir detectar qualquer sinal, os mais complexos também não serão capazes disso. Assim, é melhor começar com métodos simples de modelagem para validar a viabilidade do projeto e, depois, passar para métodos mais complexos e demorados, em vez de começar com os complexos e passar para os mais simples. Não perca meses com a construção de modelos cada vez mais complicados, esperando que talvez o próximo seja aquele que irá salvar o projeto.

### **13.1.3 O cliente acabou não querendo**

Não importa quão preciso seja um modelo ou uma análise, o que importa é que agregue valor ao stakeholder. Uma análise pode retornar informações incrivelmente interessantes para o cientista de dados, mas não para o empresário que a solicitou. Um modelo de machine learning pode fazer previsões altamente precisas, mas, se esse modelo não for implementado e utilizado, não fornecerá muito valor. Muitos projetos de ciência de dados falham mesmo depois que o trabalho de ciência de dados tenha sido realizado.

No fim de contas, uma análise, um modelo ou um dashboard de ciência de dados é um produto. Projetar e criar um produto é uma prática na qual muitas pessoas reuniram centenas de anos de pensamento coletivo. Apesar disso, todos os anos, bilhões de dólares são investidos na criação de produtos que as pessoas acabam não querendo. Desde a Nova Coca até o Google Glass, alguns produtos de alto perfil não atraem os consumidores, o mesmo ocorrendo com alguns produtos de baixo perfil. Assim como a Microsoft e a Nokia podem fazer muito esforço para criar o Windows Phone, que os clientes acabaram não comprando, um cientista de dados pode criar produtos que não são usados.

### **Exemplo de falha: previsão do valor da campanha de vendas e marketing**

Foi iniciado um projeto em uma empresa de varejo para criar um modelo de machine learning visando prever quanto retorno no investimento (ROI) as futuras campanhas publicitárias trariam. A equipe de ciência de dados decidiu construir o modelo depois de ver o quanto as equipes de marketing e vendas se esforçaram para criar planilhas de Excel que previam o valor global. Suponha que, usando machine learning e modelagem no nível do cliente, a equipe de ciência de dados tenha criado um modelo baseado em Python que previu com mais precisão o ROI das campanhas.

Mais tarde, a equipe de ciência de dados descobriu que a única razão pela qual as equipes de marketing e de vendas criaram planilhas no Excel com previsões de ROI era conseguir que o departamento financeiro os aprovasse. O financeiro recusou-se a trabalhar com outro tipo de arquivo que não fosse o Excel, pois o Python era uma caixa preta para eles. Portanto, a ferramenta não foi usada porque a equipe de ciência de dados não considerou as necessidades do cliente. A necessidade não era a previsão ser a mais precisa possível, mas, uma previsão que convenceria o financeiro de que as campanhas eram financeiramente viáveis.

A orientação universal sobre criar produtos de que os clientes irão apreciar é passar muito tempo falando e trabalhando com eles. Quanto mais entender as necessidades, desejos e problemas dos clientes, mais provavelmente desenvolverá um produto que eles desejam. As áreas de pesquisa de mercado e pesquisa de experiência de usuário são diferentes maneiras de entender o cliente, por meio de pesquisas e grupos de foco em pesquisa de mercado ou por meio de histórias de usuários, pessoas e testes em pesquisa de experiência de usuário. Muitas outras áreas criaram os próprios métodos e os têm usado há anos.

Apesar de tudo de bom que já possa ter sido pensado, a ciência de dados como área é especialmente suscetível de falhas por não entender as necessidades do cliente. Por qualquer razão, os cientistas de dados ficam muito mais confortáveis olhando tabelas e gráficos do que sair e interagir com pessoas. Muitos projetos de ciência de dados falharam porque os cientistas de dados não se esforçaram o suficiente para contatarem clientes e

stakeholders visando compreender quais eram os verdadeiros problemas. Em vez disso, os cientistas de dados saltaram para a construção de modelos interessantes e para a exploração de dados. Na verdade, essa situação é uma das principais razões pelas quais optamos por dedicar o Capítulo 12 à gestão dos stakeholders. Esperamos que já tenha uma compreensão melhor de como pensar o relacionamento com stakeholders ao ter lido aquele capítulo, mas, se o tiver ignorado, talvez convenha lê-lo.

Caso se encontre na situação de ter um produto que não pareça deslancar, a melhor coisa a fazer é falar com seus clientes. Nunca é tarde demais para falar com os clientes. Independentemente de seu cliente ser stakeholder ou cliente da empresa, a comunicação e a compreensão podem ser úteis. Se seu produto não é útil para eles, podem dizer por que não o é? Haveria como corrigir os problemas adicionando novos recursos ao produto? Talvez possa alterar uma análise adicionando um conjunto de dados diferente. É possível melhorar um modelo de machine learning ajustando o formato do resultado ou quão rapidamente é executado. Você nunca saberá até falar com as pessoas.

Também se trata do conceito de um produto mínimo viável (MVP, em inglês), que é bastante usado no desenvolvimento de software. A ideia é que quanto mais rápido conseguir um produto que funcione e que possa ser comercializado, mais rapidamente também pode obter feedback sobre o que funciona ou não e, então, repetir esse feedback. Na ciência de dados, quanto mais rápido tiver qualquer modelo funcionando ou qualquer análise feita, mais rápido pode mostrá-lo aos clientes ou stakeholders e obter feedback. Passar meses repetindo um modelo o impede de obter esse feedback.

Quanto melhor entender os clientes durante todo os processos de projeto e construção de seu trabalho, menos provável será de ocorrer uma falha de um cliente que não queira o produto. Se você acabar falhando dessa forma, a melhor maneira de avançar é começar a se comunicar para tentar encontrar uma solução.

## **13.2 Gerenciamento de risco**

Alguns projetos são mais arriscados do que outros. Pegar dados com os

quais a equipe já trabalhou e fazer um dashboard-padrão de uma maneira também padrão tem uma boa probabilidade de sucesso. Encontrar um novo conjunto de dados na empresa, construir um modelo de machine learning em torno dele, a ser executado em tempo real e exibi-lo ao cliente em uma interface de usuário agradável é um projeto mais arriscado. Como cientista de dados, você tem algum controle da quantidade de risco que assume a todo momento.

Uma grande consideração acerca dos riscos é com quantos projetos você trabalha ao mesmo tempo. Se estiver trabalhando em um único projeto arriscado, e o mesmo falhar, pode ser bastante difícil lidar com isso. Se, no entanto, puder trabalhar em vários projetos ao mesmo tempo, poderá atenuar o risco. Se um desses projetos falhar, há outros para trabalhar. Se um projeto for um modelo de machine learning extremamente complexo e que tenha uma probabilidade limitada de sucesso, você poderá trabalhar ao mesmo tempo em um simples dashboard e em um relatório; se o projeto de machine learning falhar, os stakeholders ainda poderão ficar satisfeitos com os relatórios.

Ter vários projetos também pode ser positivo do ponto de vista da utilização. Os projetos de ciência de dados têm muitas arrancadas e paradas, desde aguardar os dados até a espera do retorno dos stakeholders, além da espera que os modelos se ajustem. Se você acabar preso em um projeto por algum motivo, terá a oportunidade de progredir em outro. Pode até ajudá-lo com os bloqueios mentais; distrair-se quando está se sentindo travado pode ser uma ótima maneira de refrescar o pensamento.

Outra forma de atenuar o risco é planejar pontos de parada de um projeto. Na medida do possível, um projeto que parece que pode falhar deve ser projetado com a expectativa de que se, em um determinado ponto, ele não estiver sendo bem-sucedido, ele será interrompido. Em um projeto no qual não está claro se os dados existem, por exemplo, o escopo do projeto pode ser definido para que, se, após um mês de pesquisa, não for possível encontrar dados bons, ele seja considerado inviável e cancelado. Se a expectativa de que ele possa não funcionar for apresentada prematuramente, acabar com o projeto é menos surpreendente e menos dispendioso.

De certa forma, encerrar o projeto precocemente exemplifica o fato de que



a ciência de dados trata de pesquisa e desenvolvimento. Como a ciência de dados está repleta de incógnitas, faz sentido planejar a possibilidade de que, à medida que se aprende mais por meio do trabalho exploratório, a ideia possa não dar certo.

Embora valha a pena minimizar o risco em um portfólio de projetos, não o remova completamente. A ciência de dados tem tudo a ver com correr riscos: quase qualquer projeto suficientemente interessante terá muitas incertezas e incógnitas. Essas incógnitas arriscadas podem ocorrer porque ninguém havia usado ainda um novo conjunto de dados, ninguém em uma empresa já havia tentado determinada metodologia antes ou o stakeholder faz parte de um grupo da empresa que nunca usou a ciência de dados antes. Muitas contribuições valiosas para a ciência de dados em empresas vieram de pessoas tentando algo novo e se, como cientista de dados, tentar evitar projetos que poderiam falhar, também evitará grandes sucessos.

Embora este capítulo trate de muitas maneiras como os projetos de ciência de dados falham, as equipes de ciência de dados podem às vezes falhar na sua totalidade por não assumirem riscos suficientes. Considere uma equipe de ciência de dados que tem algumas novas ideias de projetos e relatórios, tendo êxito com elas, mas que fica estagnada, só atualizando o trabalho precedente. Embora esses projetos possam não falhar porque estão apresentando trabalho à empresa, essa equipe perderia novas áreas potenciais para a ciência de dados.

## **13.3 O que fazer quando seu projeto falhar**

Se seu projeto de ciência de dados falhar, não significa que todo o tempo que passou trabalhando nele tenha sido desperdiçado. Na Seção 13.2, descrevemos algumas ações potenciais que podem ser tomadas para superar as dificuldades enfrentadas no projeto. Mesmo que não haja uma maneira de o projeto ter êxito, ainda há passos a seguir para tirar o máximo proveito do que resta dele. Nas seções seguintes, sugerimos algumas estratégias para lidar com suas emoções quando um projeto falha.

### **13.3.1 O que fazer com o projeto**

Embora o projeto possa ter falhado, é provável que ainda haja muito a fazer para aproveitá-lo, tanto em termos de conhecimento quanto de tecnologia. Os passos a seguir podem ajudá-lo a reter muitos desses ganhos.

## **Documente as lições aprendidas**

A primeira coisa a fazer com um projeto que falhou é avaliar o que se pode aprender com ele. Algumas perguntas importantes para fazer a si mesmo e à equipe são:

- *Por que falhou?* Essa pergunta parece quase óbvia, mas, muitas vezes, é o caso de não conseguir compreender por que um projeto falhou até que volte a ter uma visão geral. Ao conversar com todas as pessoas envolvidas no projeto, é possível diagnosticar melhor o que não correu bem. A empresa Etsy popularizou o conceito de *pós-morte sem culpa* (*blameless post-mortem*) – uma discussão realizada depois de algo ter falhado, no qual uma equipe pode diagnosticar o problema sem culpar ninguém. Ao pensar em um problema como tendo sido causado por uma falha na forma como a equipe trabalha (em vez dos erros de alguém), é mais provável encontrar uma solução. Sem temer advertências, as pessoas estarão mais dispostas a falar abertamente sobre o que aconteceu.
- *O que poderia ter sido feito para evitar a falha?* Ao compreender os fatores que contribuíram para a falha, é possível entender como evitar situações similares no futuro. Se os dados não fossem suficientes para que o projeto funcionasse, por exemplo, a falha poderia ter sido evitada por uma fase exploratória mais longa. Esses tipos de lições ajudam a equipe a crescer e a amadurecer.
- *O que foi aprendido sobre os dados e o problema?* Mesmo que o projeto seja um fracasso, muitas vezes se aprendem coisas que serão valiosas no futuro. Talvez os dados não tenham um sinal, mas, para chegar a esse ponto, ainda teve de adicionar um novo conjunto de dados; agora, sim, é possível fazer essas mesmas adições com mais facilidade em outros projetos. Essas perguntas podem ajudá-lo a ter ideias de possíveis coisas que podem ser salvas do projeto e ajudá-lo a criar alternativas para o mesmo.

Ao ter uma reunião na qual a equipe trabalha com essas perguntas e, depois, salvando os resultados em um local compartilhado, agregará muito mais valor ao projeto que falhou.

## **Considere mudar a direção do projeto**

Embora o projeto em si possa ter sido um fracasso, pode haver maneiras de tirar proveito dele. Se estiver tentando criar uma ferramenta para detectar anomalias nas receitas da empresa, por exemplo, e ela falhar, talvez ainda possa usar esse mesmo modelo como uma ferramenta de previsão. Empresas inteiras foram construídas partindo de uma ideia que falhou e a ressignificando para fazer algo com êxito.

A articulação de um produto requer muita comunicação com os stakeholders e clientes. Você está basicamente de volta ao início do processo de design do produto, tentando descobrir um bom uso para seu trabalho. Ao falar com os stakeholders e os clientes, é possível compreender os problemas deles e ver se o seu trabalho é útil para algo novo.

## **Encerrar o projeto (cortar e executar)**

Se você não puder mudar a direção do projeto, a melhor coisa a fazer é encerrá-lo. Ao cancelar definitivamente o projeto, você permite a si mesmo e à equipe seguir para um trabalho novo e mais promissor. É extremamente fácil para um cientista de dados querer continuar trabalhando em um projeto para sempre, na esperança de que um dia ele funcione. (Existem milhares de algoritmos por aí; eventualmente um funcionaria, certo?) Mas se começar com um bloqueio ao tentar fazer que algo funcione, acabará investindo esforço desnecessário. Além disso, não é divertido trabalhar na mesma coisa até o fim dos tempos! Embora cortar um projeto seja difícil, pois ele exige que você admita que ele não vale mais o esforço, ele o compensará a longo prazo.

## **Comunique-se com seus stakeholders**

Um cientista de dados deve se comunicar com os stakeholders ao longo de um projeto de ciência de dados (consulte o Capítulo 12), mas a comunicação deve ser maior se o projeto estiver apresentando falhas.

Embora possa se sentir confortável em esconder riscos e problemas dos stakeholders para evitar desapontá-los, acabar em uma situação na qual um stakeholder é surpreendido ao saber que o projeto falhou pode ser catastrófico para uma carreira. Ao deixar os stakeholders saberem que problemas estão ocorrendo ou que o projeto não pode mais avançar, você está sendo transparente com os stakeholders e inspirando confiança. Depois de ajudá-los a entender o estado do projeto, vocês podem trabalhar juntos para decidir as próximas etapas.

Se não tiver certeza sobre como comunicar os problemas com um stakeholder, seu gerente deve ser um bom recurso. Eles podem fazer um brainstorming de uma abordagem para passar a mensagem ou até assumir a liderança em entregá-las eles mesmos. Pessoas e empresas diferentes gostam de ter mensagens transmitidas de diferentes formas, desde planilhas que mostram os problemas com codificação de cores verde/amarela/vermelha até conversas tomando um café. Seu gerente ou outras pessoas da sua equipe devem saber o que funciona melhor.

É comum que você, como cientista de dados, sinta-se ansioso ao comunicar que um projeto está apresentando falhas; você se sente emocionalmente vulnerável e pensa que está em uma posição de fraqueza. Embora haja ocasiões nas quais a notícia é mal recebida, outras pessoas estão muitas vezes dispostas a ajudá-lo a trabalhar para resolver problemas e decidir as etapas seguintes. Após comunicar a falha do projeto, você pode se sentir aliviado e não ficar sofrendo.

### **13.3.2 Lidando com emoções negativas**

Esqueça-se um pouco do projeto e da empresa: você também precisa pensar no seu próprio bem-estar. Ter um projeto que falha é emocionalmente difícil e a pior coisa! Se não for cuidadoso, um projeto que falhou pode sugar suas energias e assombrá-lo por muito tempo depois do término desse. Sendo cuidadoso consigo mesmo sobre como reage à falha e à história que constrói sobre ela, é possível se preparar para ter sucesso mais adiante.

Um monólogo interno natural no fim de um projeto que falhou é: “Se eu fosse um cientista de dados melhor, o projeto não teria falhado”. Esse pensamento é uma falácia: a maioria dos projetos de ciência de dados falha

porque a ciência de dados baseia-se em coisas que nunca poderiam funcionar. Grande parte dos grandes cientistas de dados esteve envolvida com, ou até mesmo liderou, projetos que não foram bem-sucedidos. Ao colocar em si mesmo a culpa pela falha do projeto e por possíveis deficiências de ciência de dados, você está colocando o peso de todo o projeto em si mesmo. Conforme já discutido neste capítulo, há muitas razões pelas quais os projetos de ciência de dados falham, e é muito raro que a questão seja a competência do cientista de dados. É muito comum ficar ansioso pensando que o projeto esteja apresentando falhas por sua causa, mas essa ansiedade está em sua cabeça e não é um reflexo da realidade.





Caso se permita falhar e aceitar que a falha não é um sinal de fraqueza, você será mais capaz de aprender com a experiência. Ter confiança em si mesmo e nas suas competências facilita pensar na falha e no que contribuiu para ela, porque não vai machucar tanto. Dito isso, a capacidade de ter confiança e de defender uma falha é algo que leva tempo, paciência e prática, por isso não se surpreenda se estiver se debatendo para ter mais confiança. Tudo bem!

O ponto-chave aqui é que a melhor coisa a fazer por si mesmo quando um projeto falha é entender que o fracasso não é um reflexo das suas competências. Os projetos falham por motivos fora do controle, e você conseguirá superar a falha. Quanto mais conseguir entender isso, mais fácil será aceitar a falha.

Terminaremos este capítulo com uma metáfora para a ciência de dados. É comum que os aspirantes e cientistas de dados iniciantes pensem em um cientista de dados profissional como sendo um arquiteto de edifícios. Um arquiteto principiante pode projetar casas simples e um arquiteto experiente pode construir arranha-céus, mas, se um deles tiver um prédio colapsado, isso é um fracasso profissional. Da mesma forma, uma maneira de ver um cientista de dados é que ele constrói modelos cada vez mais complexos, mas, se falhar, a carreira está comprometida. Depois de ler este capítulo, esperamos que reconheça que *este não é um modelo correto de um cientista de dados profissional*.

Uma metáfora melhor: um cientista de dados é como um caçador de

tesouros (Figura 13.1) que procura objetos de valor perdidos e, com sorte, encontrará alguns! Um caçador de tesouros principiante pode procurar bens comuns, enquanto o experiente encontra o tesouro mais lendário. Um cientista de dados se parece mais a um caçador de tesouros; ele procura modelos de sucesso, e, às vezes, seus modelos e análises funcionam! Embora um cientista de dados sênior possa trabalhar em projetos mais complexos, todo mundo sempre falha, e isso é apenas parte do trabalho.

<input type="checkbox"/> Modelo de arquiteto <input type="checkbox"/>		<input checked="" type="checkbox"/> Modelo de caça ao tesouro <input checked="" type="checkbox"/>	
			
Cientista de dados júnior	Cientista de dados sênior	Cientista de dados júnior	Cientista de dados sênior

*Figura 13.1 – Duas metáforas para a ciência de dados: arquitetura e caça ao tesouro.*

### 13.4 Entrevista com Michelle Keim, chefe da equipe de ciência de dados e machine learning da Pluralsight

Michelle Keim lidera a equipe de ciência de dados e machine learning da Pluralsight, uma plataforma de aprendizagem de tecnologia corporativa com a missão de democratizar as competências tecnológicas. Tendo criado e liderado equipes de ciência de dados em várias empresas, incluindo a Boeing, a T-Mobile e a Bridgepoint Education, ela compreende muito bem a razão pela qual os projetos de ciência de dados podem falhar e como lidar com isso.

#### Quando foi que você passou por um fracasso em sua carreira?

Liderei um projeto para construir um conjunto de modelos de retenção de

clientes. Pensei que tinha falado com todos os stakeholders certos e que havia compreendido a necessidade da empresa, como a equipe trabalhava e a razão pela qual os modelos eram necessários. Construímos os modelos, mas logo percebemos que não havia interesse nenhum neles. O problema foi que não conversamos com os agentes de atendimento ao cliente que realmente usariam o resultado; tratamos do assunto com os líderes. Apresentamos uma lista de probabilidades de que um cliente iria embora, mas os agentes de atendimento não sabiam como lidar com isso. Eles precisavam saber o que deveriam fazer quando corressem o risco de um cliente ir embora, o que é um problema muito diferente daquele que tínhamos abordado. A maior lição aprendida foi que realmente é preciso começar desde a raiz e entender o caso de uso do problema. Qual é o problema que será resolvido pelas pessoas que usarão o resultado?

### **Há sinais de alerta visíveis antes do início de um projeto?**

Penso que, em parte, é uma intuição que se adquire com a experiência. Quanto mais coisas você vê indo mal e quanto mais aproveitar a oportunidade para aprender com as falhas, mais fácil será identificar os sinais de alerta. A chave é manter seu ciclo curto, para ter a oportunidade de detectá-los o mais precocemente possível; é preciso receber feedback com frequência.

Os cientistas de dados tendem a ficar entusiasmados com o trabalho e a se esquecerem de levantar a cabeça. É realmente importante não apenas compreender aonde você quer ir ao fim do dia, mas também como o sucesso é encontrável em diferentes pontos ao longo do caminho. Assim, é possível comparar seu trabalho com isso, receber feedback e ser capaz de mudar a direção, se necessário. Os checkpoints (pontos de verificação) permitem saber rapidamente quando perdeu ou interpretou mal alguma coisa e corrigi-la, em vez de somente percebê-la no final e ter de recuar.

### **Como uma falha é tratada em diferentes empresas?**

A maneira de tratar falhas está altamente atrelada à cultura da empresa. Aconselharia as pessoas que estão em busca de emprego a tentarem saber se a empresa tem uma cultura de aprendizado e feedback constante. Na

entrevista, há a oportunidade de perguntar ao entrevistador: “O que você está aprendendo por conta própria?”, “Como surgiu essa oportunidade para você?”, “Se eu assumisse a função, como receberia feedback?”, “É algo que se precisa buscar ou é formalizado?” Sentir como os funcionários respondem a essas perguntas pode evidenciar muita coisa.

Quando você já está em uma empresa, há perguntas que pode tentar responder por si mesmo para ver se há uma cultura saudável. Depois que um projeto é encerrado, há alguma oportunidade de pausar e olhar para trás? Você tenta aprender retrospectivamente no fim dos projetos? Você vê a liderança manter uma comunicação aberta e assumir a responsabilidade pelas falhas em vários níveis na empresa? Você sente medo também quando não existe uma cultura forte. Começa a ver os comportamentos que são mais para benefício próprio do que para servir aos outros, e esse tipo de comportamento não é saudável.

### **Como saber se um projeto em que está apresenta falhas?**

Você não pode saber se está falhando se não definiu que é o sucesso desde o início. Quais são os objetivos que está tentando alcançar e como os checkpoints (pontos de verificação) ao longo do caminho para o sucesso se parecem? Se desconhece isso, está apenas tateando no escuro para descobrir se o projeto está indo bem ou não. Para se preparar para o sucesso, é preciso certificar-se de que colaborou com os stakeholders para obter uma resposta bem definida para tais questões. Você precisa saber por que está participando desse projeto e qual problema está tentando resolver ou não saberá o valor do que está apresentando e se sua abordagem está correta. Parte da função de um cientista de dados é trazer sua experiência ao jogo, ajudar a enquadrar o problema e definir as métricas de sucesso.

### **Como pode superar o medo de falhar?**

É preciso lembrar-se de que algumas falhas são necessárias, porque, se tudo correr perfeitamente, nunca aprenderia nada. Como cresceria? Essas experiências são necessárias, pois não há substituição para lidar com falhas e ficar bem com elas. É verdade que o fracasso pode ser doloroso e você pode se perguntar: “E agora, o que farei?”. Mas depois de entrar nos eixos,



aprender com isso e se voltar para o próximo projeto, essa resiliência transforma-se em confiança. Se você souber que algumas coisas podem dar errado, tudo será mais fácil na próxima vez. Além disso, se garantir que irá receber feedbacks frequentes, conseguirá identificar as falhas antes que se tornem devastadoras. Ninguém espera perfeição! O que se espera é que seja honesto sobre o que não sabe e que continue aprendendo a fazer perguntas e a procurar feedback.

## **Resumo**

- Os projetos de ciência de dados geralmente falham por causa de dados inadequados, falta de sinal ou por não serem adequados ao cliente.
- Depois que um projeto falhar, catalogue o motivo e considere mudar de direção ou encerrá-lo.
- Uma falha em projeto não é um reflexo da qualidade do cientista de dados.
- Um cientista de dados não é o único responsável pela falha do projeto.

## CAPÍTULO 14

# Como participar da comunidade de ciência de dados

Este capítulo abrange:

- Como aumentar seu portfólio de projetos e posts de blog
- Como encontrar e tirar o máximo de proveito de conferências
- Como proferir uma ótima palestra de ciência de dados em um meetup ou conferência
- Como contribuir para o código aberto

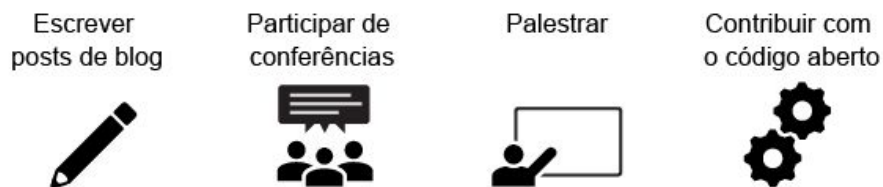
Em ciência dos dados parece que a única maneira de prosperar na carreira é fazendo um bom trabalho. Mas há muitas outras maneiras de melhorar suas competências, especialmente envolvendo-se com a comunidade dessa área. Passar tempo fora do trabalho em atividades como proferir palestras ou contribuir para o código aberto pode ser extremamente benéfico para sua carreira.

Neste capítulo, falaremos sobre quatro maneiras de participar da comunidade: aumentando seu portfólio, participando de conferências, palestrando e contribuindo com o código aberto. Apresentamos quatro atividades para que possa escolher aquelas de que você gosta mais, uma vez que muito poucas pessoas têm tempo e energia para participar de todas. Embora demandem algum tempo fora do seu trabalho diário normal, não significa que devem tomar todo o tempo da sua vida. Aqui, damos conselhos sobre como usar seu tempo com eficácia e táticas, tais como reutilizar palestras, transformar um post de blog em uma palestra e escrever um post de blog sobre sua primeira contribuição ao código aberto.

Embora essas atividades possam ser úteis e imensamente gratificantes, não é necessário participar delas para ter uma carreira de sucesso na ciência

de dados. Muitos cientistas de dados, incluindo aqueles que estão em cargos que requerem mais experiência, não participam dessas atividades. Mas nós, as autoras, sentimos que fazer parte da comunidade nos ajudou muitas vezes em nossas carreiras, incluindo o recebimento de ofertas de emprego e as promoções obtidas. O trabalho público é uma área na qual o tempo investido pode ser recompensado em dobro.

São apresentados quatro benefícios principais de participar da comunidade mais ampla de ciência de dados (Figura 14.1):



*Figura 14.1 – Algumas maneiras de participar da comunidade que são abordadas neste capítulo.*

- *Como aprender competências* – ao se envolver com a comunidade, você aprende novas técnicas, às quais não seriam apresentadas caso dependessem do seu trabalho diário. A criação de um projeto de código aberto é a atividade que mais diretamente desenvolve suas competências técnicas, pois escreverá código para que outras pessoas usem e trabalhem colaborativamente em um projeto técnico. Todavia, cada atividade tem seus benefícios. Escrever textos em blog é uma ótima maneira de abordar lacunas do seu conhecimento e receber feedback. Palestrar ajuda a aperfeiçoar suas competências de apresentação, o que pode auxiliá-lo a convencer um stakeholder de que você precisa de aportes financeiros ou de que ele deve apoiar o seu projeto. Assistir a uma palestra certa em uma conferência pode desbloquear um projeto importante e poupar-lhe horas de trabalho.
- *Como aumentar sua rede de contatos* – conectar-se com a comunidade é uma ótima maneira de encontrar um grupo de colegas que entendem suas dificuldades. Mesmo que tenha colegas na empresa, pode ser que esses não sejam especialistas em um determinado nicho, mas você pode contar com o conselho de um membro da comunidade. Igualmente você pode aprender como é trabalhar na ciência de dados em diferentes

empresas.

- *Como conseguir oportunidades* – quanto mais estiver envolvido na comunidade, mais será solicitado para auxiliar em projetos, palestrar ou falar em um podcast, inclusive com possibilidade de conseguir seu próximo emprego graças a alguém que o encontrou pelo seu trabalho online ou depois de conhecê-lo em uma conferência. Trata-se de um grande ciclo de feedback positivo: as palestras levam a mais palestras e projetos levam a mais projetos. Essas oportunidades podem ser informativas, interessantes e divertidas.
- *Retribuir* – este é o benefício menos direto para você, mas fundamental ao bem-estar da comunidade. Quando perguntar a um mentor sobre como retribuir o apoio que recebeu, muitos lhe dirão: “Passe adiante. Ajude outros e torne-se um mentor para eles”. Ser parte da comunidade pode tornar o trabalho de ciência de dados muito mais gratificante. Ao realizar tarefas que ajudam os outros, você reconhecerá seu próprio valor, como se estivesse trabalhando por muito mais do que apenas por um salário.

## 14.1 Como aumentar seu portfólio

Só porque você tem um trabalho agora não significa que pode esquecer todos os excelentes hábitos que você desenvolveu para conseguir o emprego. No Capítulo 4, você aprendeu a escrever posts de blog e a construir um portfólio. Estar empregado não significa que não haja valor em continuar escrevendo, mantendo essa atividade e a expandindo. Trabalhar em um blog ou em um projeto paralelo não precisa ser difícil. Neste capítulo, discutimos novos tópicos e maneiras de reciclar seu trabalho para utilizá-los com o mesmo entusiasmo.

### 14.1.1 Mais posts de blog

Esperamos que esteja aprendendo muitas coisas novas em seu trabalho como cientista de dados. Assim, como otimizar consultas SQL em uma tabela com 30 bilhões de linhas? Como trabalhar de maneira eficaz com a equipe de marketing? Quais estratégias usar para começar a navegar por

centenas de tabelas?

Se estiver em uma empresa que emprega outros cientistas de dados, aprenderá diretamente com esses profissionais, seja lendo o código deles ou programando em pares. É uma boa ideia tomar notas durante esse processo, pois receberá muitas informações novas e é pouco provável que se lembre de tudo em alguns meses. Quando está fazendo isso, por que não compartilhar suas notas com a turma (neste caso, toda a internet)? Desconhecidos na internet não são os únicos que se beneficiarão com aquilo que compartilha e escrever um post de blog é uma ótima maneira de consolidar sua aprendizagem. Você pode até mesmo encontrar-se com essas pessoas anos mais tarde quando se referir a um tutorial que escreveu anos antes.

Se seguiu nosso conselho no Capítulo 4, você já deve ter um blog com alguns posts. Se não tiver feito isso ainda, mas estiver interessado em começar um, recomendamos voltar àquele capítulo e seguir os passos dele. Tudo o que escrevemos lá ainda se aplica; as mesmas estratégias para posts eficazes de blog quando você estava procurando seu primeiro trabalho de ciência de dados se mantêm até mesmo depois de estar trabalhando na área. A única grande mudança é que você precisa ter certeza, se estiver escrevendo sobre projetos que fez no trabalho (em comparação a competências gerais aprendidas de programação, estatística ou gestão de pessoas), de que não está compartilhando informações confidenciais ou proprietárias e de seguir quaisquer outras regras que sua empresa tenha para os blogs pessoais dos funcionários (como enviar os posts para o departamento de comunicações primeiro).

Se não quiser ter seu próprio blog, veja se sua empresa tem um blog técnico. Mesmo que posts anteriores tenham sido centrados em engenharia, você pode fazer um post de ciência de dados. Pode levar algum tempo para receber aprovação, mas um bônus de seguir essa rota é poder escrever seus pensamentos durante o trabalho. Mesmo que sua empresa não tenha um blog público, ela deve ter alguma documentação interna e treinamento. Se você perceber que teve de aprender algo perguntando a várias pessoas ou vasculhando instruções desatualizadas, crie ou atualize um material com instruções claras para novos funcionários. Se o tópico for algo que será útil

para alguém de fora da empresa (não em uma ferramenta interna proprietária ou descrição de dados, por exemplo), você pode mais tarde transformá-lo em um post de blog ou palestra.

### **14.1.2 Mais projetos**

Os projetos de ciência de dados (que também tratamos no Capítulo 4) são aqueles nos quais ou se escolhe ou cria um conjunto de dados e analisa-o para responder a uma questão. Você poderia usar a API do Twitter para fazer uma análise de rede dos usuários tuitando sobre uma conferência de ciência de dados, por exemplo. Em alguns casos, um projeto sequer precisa ser uma análise; talvez você exiba suas competências de engenharia criando um bot no Slack para permitir que os usuários se deem “pontos”, mantendo o controle dos totais em um banco de dados que você configurou.

Os projetos podem ser muito mais difíceis do que manter um blog atualizado. Dependendo do setor, sua empresa pode estar aberta ou até mesmo incentivar a escrita de textos longos sobre o seu trabalho. Mesmo que não seja, é possível escrever posts não técnicos sobre como lidar com stakeholders ou a respeito de sua experiência no mercado de trabalho. Poucas empresas, no entanto, compartilham dados publicamente; portanto, mesmo que pudesse compartilhar o código de uma análise excepcional que tenha feito, isso não teria muito sentido, pois não poderia compartilhar os dados ou os resultados. Se quiser compartilhar as análises que fizer, terá de fazê-lo como um projeto paralelo no seu tempo livre.

Dito isso, é bom fazer um projeto paralelo de vez em quando. Por um lado, quando você deseja mudar de emprego, as empresas podem pedir um exemplo de uma análise de dados que tenha feito. Se trabalha na sua empresa há alguns anos, não mostre um projeto que foi feito em um bootcamp ou quando se inscreveu em um MOOC, mas, sim, mostre como suas habilidades evoluíram durante o trabalho como cientista de dados. Os princípios para encontrar um tópico e escrever uma boa análise permanecem os mesmos do Capítulo 4.

A boa notícia é que um projeto não precisa levar muito tempo. David Robinson, que entrevistamos no Capítulo 4, faz um screencast semanal no qual ele se grava realizando uma análise em um conjunto de dados que

nunca viu antes (do projeto de Tidy Tuesday <https://github.com/rfordatascience/tidytuesday/>). Essa análise leva cerca de uma hora para ser feita, pois ele não prepara nada, mas, quando o código é carregado no GitHub, poderia servir como um exemplo de projeto de análise. É claro que apenas um cientista de dados bastante experiente faria uma boa análise tão rapidamente; todavia, qualquer um pode tentar definir um limite de tempo para a análise a fim de ajudar a manter-se concentrado no compartilhamento de resultados (e não gastar 14 horas em um projeto que nunca é mostrado a alguém).

## 14.2 Participação em conferências

Às vezes, fazer parte da comunidade significa ter de sair de casa. Na maior parte do tempo, significa ir a conferências, nas quais as pessoas que estão (ou querem estar) em setores ou áreas similares reúnem-se para falar sobre o próprio trabalho.

As conferências geralmente são eventos anuais que ocorrem no país e no exterior. A área da ciência de dados realiza muitas conferências. Strata, rstudio::conf, PyData, EARL e Open Data Science Conference são apenas alguns dos eventos mais importantes, sendo que muitos deles têm ramificações regionais. Você também pode se interessar por conferências de tecnologia mais gerais que se sobrepõem à ciência de dados, como Write/Speak/Code, PyCon, Grace Hopper e SciPy.

As conferências tendem a durar de dois a quatro dias, com programação desde o início da manhã até a noite, além de atividades sociais posteriores. Podem ser de um único eixo (apenas uma palestra acontecendo em um determinado momento) ou de vários eixos, mas todas contam com vários palestrantes. Também podem ser muito caras – normalmente a um custo diário entre 300 e 700 dólares apenas para o ingresso, isso nos EUA. Algumas conferências também têm workshops de meio a dois dias por uma taxa extra, normalmente de cerca de 750 dólares por dia.

Uma razão pela qual estamos informando preços aproximados é porque há muitas maneiras de pagar menos do que o preço total do ingresso. Se for membro de alguma minoria, procure bolsas de estudo ou códigos de

desconto oferecidos a todos os membros de um grupo, como o R-Ladies ou o PyLadies. Se trabalhar em uma organização sem fins lucrativos ou na área acadêmica, também é possível pagar menos, além de muitas das grandes conferências oferecem descontos na compra antecipada. Outra ótima maneira de reduzir as despesas é ser palestrante, pois talvez o ingresso seja oferecido gratuitamente. Por fim, algumas conferências disponibilizam bolsas às quais pode se candidatar e que cobrirão o custo do ingresso e talvez o custo total, incluindo transporte e acomodação.

Tendo em conta o potencial preço elevado, por que razão deve investir seu tempo e dinheiro (ou o dinheiro do seu empregador) para participar, especialmente se a conferência for gravada e as palestras disponibilizadas online? Aqui, retornamos a um dos benefícios principais que falamos no início do capítulo: a rede de contatos. A rede contatos pode ter uma conotação negativa: alguém que cumprimenta a todos em uma sala, na esperança de encontrar uma pessoa importante e de lucrar algo com isso. Mas a rede de contatos preocupa-se mais em encontrar uma comunidade de pessoas que o apoiem. Esse apoio pode vir de formas muito tangíveis, como alguém que o apresenta a uma pessoa que trabalha em uma empresa à qual está se candidatando, ou intangível, como a sensação de estar finalmente em um local de pessoas da tecnologia, na qual a grande maioria é formada por mulheres.

A formação de uma rede de contatos é melhor ser feita a longo prazo. Por isso, mesmo que não tenha ninguém com o que você acha que precisa de ajuda no momento, é ótimo montar essa base antes de procurar um novo emprego ou um parceiro para um projeto de código aberto.

**CÓDIGO DE VESTUÁRIO** Uma pergunta comum para os participantes de primeira viagem é o que vestir. Em geral, as conferências são informais. Para uma conferência específica, veja se consegue encontrar imagens da conferência no Twitter ou no site do evento. Uma coisa para se ter em mente, no entanto, é que os palestrantes podem estar vestidos de maneira mais formal do que o público; só porque todos os palestrantes estão vestidos de maneira um pouco mais formal não significa que o público também tenha que se vestir assim. Se ficar muito em dúvida, leve roupas que sejam um meio-termo, como um vestido ou uma polo e uma calça de brim escura. É raro que uma conferência seja tão formal para usar terno ou, então, bermuda e camiseta. Geralmente, haverá pelo menos alguma escala de vestuário das pessoas; portanto, é improvável que se sinta um peixe fora d'água!

Com tantas conferências acontecendo, como descobrir quais são as que



valem a pena participar? As conferências variam, mas aqui estão alguns eixos a considerar:

- *Acadêmico* – algumas conferências, como useR!, NeurIPS e JSM, têm um grande número de participantes da área acadêmica ou em empregos de pesquisa profunda. No outro extremo, há algumas conferências em que essencialmente todos são estudantes ou professores. Se estiver no setor, talvez não ache tantas palestras aplicáveis; embora as pessoas no setor possam indicar palestras, as apresentações podem ser um algoritmo de machine learning de ponta, útil somente se trabalhar em uma empresa gigante de comércio eletrônico.
- *Tamanho* – as conferências podem contar com um número variável de 150 a dezenas de milhares de pessoas. Recomendamos que comece com uma conferência de pequena à média, entre 200 e 1.500 participantes. O tamanho menor significa que é menos intimidante navegar e é mais provável encontrar as mesmas pessoas várias vezes, levando a conexões mais fortes.
- *Empresas em contratação* – por outro lado, você pode querer participar de uma grande conferência porque está procurando emprego. Embora possa encontrar alguém que esteja contratando em uma conferência, algumas de maior porte têm feiras de emprego, onde os empregadores pagam especificamente por um estande para conversar com funcionários em potencial.
- *Nível de palestras* – a maioria das conferências é geralmente para pessoas que estão trabalhando ou estudando na área. Se você não souber a linguagem R, por exemplo, provavelmente não aproveitaria muito o rstudio::conf, uma conferência dirigida pela empresa que desenvolve o ambiente de desenvolvimento interativo (IDE) principal para R. As palestras da conferência são em geral destinadas a um nível intermediário de conhecimento geral, mas variam no nível de conhecimento específico esperado. A rstudio::conf, por exemplo, poderia organizar uma palestra introduzindo um pacote para séries cronológicas. Seria preciso ter algum conhecimento de R para entender a palestra, mas o palestrante não espera que tenha muita experiência

trabalhando com séries cronológicas. Ou uma palestra em uma conferência sobre experimentação online pode ser uma introdução sobre como a pesquisa qualitativa pode complementar os métodos quantitativos.

- *Diversidade e inclusão* – infelizmente, nem todos os organizadores estão preocupados em garantir que as conferências sejam acolhedoras para todos. Se constatar que todos os 45 palestrantes são homens, dá para assumir que boa parte dos participantes também será. Além da escalação de palestrantes, olhe o site para saber se há um código de conduta. Se precisar de determinadas acomodações, como um local acessível para cadeiras de rodas, procure um endereço de email no site da conferência e envie um email para perguntar.
- *Especialidade* – assim como a ciência de dados tem muitas especialidades, as conferências também. Quer esteja à procura de uma linguagem ou domínio específico, provavelmente há uma conferência destinada para você.

Quando decidir o tipo de conferência em que está interessado, procure avaliações das conferências antes de se comprometer a ir a uma. Se não conhecer alguém pessoalmente que assistiu a uma, pergunte no Twitter ou no LinkedIn. Observe também a programação da conferência ou, se ainda não estiver disponível, consulte a programação do ano anterior. Se houver gravações de uma palestra, assista a algumas. É melhor certificar-se de que seu investimento vale a pena. Infelizmente, algumas conferências não têm palestrantes muito bons.

Participar dessas conferências pode ser útil para sua carreira e também para seu empregador, que o tem como representante da empresa aprendendo coisas que podem ajudá-lo a trabalhar melhor. Como resultado, você pode conseguir persuadir sua empresa a pagar a conferência na totalidade ou em parte. Algumas empresas têm um orçamento formal de conferência ou de treinamento, o que é ótimo, porque o dinheiro é destinado para seu propósito, mas, se você quiser exceder esse orçamento, será difícil conseguir que seja aberta uma exceção.

Além do custo, a maioria das conferências acontece, pelo menos parcialmente, durante a semana; portanto, é preciso tirar um ou dois dias de

folga. Mas você não quer que lhe sejam descontados dos dias de férias, certo? Você precisa argumentar com seu gerente que vale a pena passar o dia na conferência em vez de trabalhar normalmente. Algumas empresas têm uma política, que pode ou não ser formal, sobre quantos dias de conferência você pode pedir. Nas empresas de tecnologia, é bastante normal assistir a pelo menos uma conferência, mas outros setores podem não adotar esse sistema.

Se precisar de argumentos para seu gerente, a seguir apresentam-se alguns benefícios para se concentrar:

- *Recrutamento* – custa milhares ou mesmo dezenas de milhares de dólares recrutar um cientista de dados. Um dos maiores problemas é, em primeiro lugar, conseguir bons candidatos. Grandes empresas de tecnologia, como a Google e a Amazon e as startups mais famosas, podem ter ótimos candidatos batendo na porta, mas a maioria das empresas não tem esse reconhecimento de nomes. Se conhecer pessoas nas conferências, você fará propaganda da sua empresa. Essa publicidade é bastante ampliada se ministrar uma palestra, tópico abordado na Seção 14.3.
- *Conhecimento* – seu gerente quer saber o que poderá fazer depois de uma conferência que não conseguiria fazer antes. É ainda melhor se compartilhar esse conhecimento com a equipe escrevendo um artigo (que também pode se tornar um post de blog!) ou apresentação. Comece observando a programação da conferência e informe ao seu gerente as palestras que serão imediatamente aplicáveis para resolver seus problemas. Lembre-se de que as conferências também têm corredores: conversas informais acontecem fora das apresentações. Talvez encontre alguém ali a solução para um problema que está enfrentando! Cinco ou dez minutos do tempo da pessoa certa podem pagar o custo do ingresso.

Se residir em uma cidade grande ou perto dela, procure lá uma conferência para não ter de pagar por viagens ou acomodações. No geral, a melhor estratégia ao solicitar dinheiro da empresa é mostrar como o que aprende com a conferência irá ajudá-lo a realizar ganhos ou a ter impactos na empresa.

### **14.2.1 Como lidar com a ansiedade social**

É um clichê que engenheiros e cientistas são introvertidos, mas a maioria das pessoas luta com a ansiedade social em algum momento. Até indivíduos mais confiantes não entram em uma sala cheia de estranhos ou sentem-se totalmente confortáveis. O que fazer se você ficar tão nervoso a ponto de ir a uma conferência e ficar metido em um canto olhando seu telefone o tempo inteiro?

Felizmente, a vantagem de assistir a uma conferência é que há algo para se falar! Em geral, uma boa estratégia é fazer perguntas; as pessoas gostam de falar sobre si mesmas. Você pode perguntar por que estão na palestra, há quanto tempo têm programado na linguagem X ou se já estiveram nessa conferência antes. Lembre-se de que muitas delas estão se sentindo desajeitadas, não só você. Se estiver nervoso, um momento agradável para tentar falar com as pessoas é nos poucos minutos que antecedem a palestra, quando você estiver sentado. Sente-se ao lado de alguém e comece uma conversa. Se seus medos de uma conversa desagradável forem concretizados, você sabe que a conversa pode durar somente alguns minutos, pois a palestra está quase começando!

Quando estiver em uma sala com muita gente, procure aqueles que estão no formato do Pac-Man: um círculo com uma abertura. Dirija-se até essa abertura, reorganize e tente fazer outra abertura para que mais pessoas se juntem. Não é preciso se apresentar assim que chegar; pode esperar por uma pausa na conversa ou até mesmo participar da conversa sem se apresentar, sobretudo se o grupo for grande.

Falamos sobre a síndrome do impostor no Capítulo 8, e essa é outra área em que ela pode atacar. Talvez tenha assistido a uma palestra que mexeu muito com você. O mais importante a se lembrar é que não deve se sentir um impostor. Se o tratarem como se você não valesse o tempo delas ou se não souber um termo ou se fizerem um comentário depreciativo, é problema delas. Algumas outras conferências serão acolhedoras. Muitas pessoas adoram ajudar os outros e recordam como era ser novato na área. Se você participar de uma conferência da qual não gosta, tente não deixar essa experiência dissuadi-lo a não participar outra vez.

Embora tenhamos abordado neste capítulo algumas estratégias para

conhecer pessoas, é totalmente normal tomar um tempo para si mesmo durante a conferência. Vários dias de socialização com desconhecidos podem ser exaustivos. É um erro comum sentir que precisa fazer algo produtivo em todos os momentos da conferência, seja assistindo a uma palestra ou formando uma rede de contatos. Mas você absolutamente não precisa! Não se sintam mal em fazer uma caminhada sozinho em vez de assistir a uma palestra durante uma sessão; você aproveitará mais da conferência se tiver um tempo para se recarregar.

## 14.3 Como ministrar palestras

Ministrar palestras pode oferecer muitas oportunidades de crescimento e abrir espaço para ir a mais palestras e conferências. Uma dificuldade que você pode enfrentar é conseguir tempo o suficiente para ir aos eventos para melhorar suas competências e rede de contatos, mas ministrar palestras é uma ótima maneira de representar a sua empresa (além dos benefícios monetários, o que torna menos oneroso ao seu empregador). Embora possa parecer que é preciso ser um especialista no setor, um palestrante espetacular ou alguém muito sociável para ministrar uma palestra, não é esse o caso. Ministrar palestras é, na verdade, uma grande estratégia para alguém introvertido. Após sua palestra, as pessoas virão até você para elogiá-lo, fazer perguntas ou apenas se apresentarem. Fazer com que sua palestra seja um tópico de conversa é uma versão melhorada da vantagem geral de falar sobre o tema da conferência.

Esta seção poderia ser um livro inteiro e, na verdade, na seção de recursos para os capítulos 13–16, recomendamos um livro sobre falar em público. Queremos enfatizar que os requisitos para ministrar uma boa palestra são menores do que se imagina. Não se trata de uma palestra do TED ou de uma grande conferência. Essas pessoas têm muita experiência e provavelmente contrataram um coach para treiná-las para falar em público. Acreditamos que se quiser ministrar uma boa palestra, você deve se concentrar em duas coisas: entreter as pessoas e motivá-las. Se as pessoas não estiverem interessadas enquanto estiver falando, será muito difícil ensinar-lhes qualquer coisa. Além disso, as pessoas retêm pouco de uma

palestra de 20, 30 ou até 60 minutos. Mas se instigar o desejo de aprender mais e equipar o público com as ferramentas para começar a fazer isso, já terá agregado muito valor.

### **14.3.1 Como conseguir uma oportunidade**

Como encontrar oportunidades para ministrar palestras? O melhor lugar para começar é procurar conferências que tenham chamadas para apresentação. Você pode enviar um breve *resumo* da sua palestra, e os organizadores escolherão palestrantes entre os que enviaram resumos. Algumas conferências fazem revisões cegas, isto é, escolhem resumos sem saber nada sobre os palestrantes, embora outros queiram saber mais sobre você.

Quando estiver procurando conferências para palestrar, aplique critérios similares aos usados quando estiver procurando conferências. Se ir a uma conferência com 10 mil pessoas parece um pesadelo, provavelmente não irá querer palestrar nela. Além disso, palestrar é uma ótima maneira de reduzir o custo de participação em um evento, então, por que não se beneficiar disso em uma conferência na qual realmente queira ouvir as demais palestras? Pergunte às pessoas que conheceu online ou em meetups quais conferências recomendam; nem sempre é fácil encontrar conferências menores.

A primeira parte de um bom resumo está em prestar atenção ao que a conferência está pedindo. Mesmo que escreva o melhor resumo de 500 palavras, você não será aprovado se os organizadores estavam pedindo um de 150 palavras. O mesmo se aplica se enviar o resumo para uma palestra sobre engenharia de dados a uma conferência que se concentra em estatística.

No geral, um bom resumo tem uma primeira frase que serve de gancho: atrai o público para aprender mais. Então, você deve explicar o problema que está resolvendo e dar uma visão geral do que o público aprenderá. Eis um exemplo de uma das palestras da Jacqueline:

*O conceito de deep learning parece complicado e difícil, mas não é exatamente assim. Graças a pacotes como o Keras, você pode começar com apenas algumas linhas de código. Uma vez que compreenda os*

*conceitos básicos, poderá usar o deep learning para fazer piadas geradas por IA! Nesta palestra, farei uma introdução ao deep learning, mostrando como pode usá-lo para criar um modelo que gera nomes de animais de estimação estranhos, como Shurper, Tunkin Pike e Jack Odins. Se souber como fazer uma regressão linear, você pode entender como criar projetos divertidos com deep learning.*

Quando tiver uma ideia abstrata, uma ótima maneira de começar é pensar em quem você era três meses, seis meses ou um ano atrás. O que você sabe agora que queria saber na época? É fácil imaginar que todo o mundo já conhece essas coisas, mas até mesmo aquilo considerado básico, como utilizar o git e o GitHub, ou como fazer web scraping, milhares de pessoas por aí não sabem, mas se beneficiariam disso. Também é possível escolher explorar sua subárea, se for interessante a um público maior. Talvez você possa ensinar a fazer mapas interativos, usar um pacote personalizado para análise rápida de dados ou explicar o que é um modelo linear generalizado. Você não precisa ser um especialista na área; na verdade, as pessoas que acabaram de aprender alguma coisa são muitas vezes os melhores professores. Aqueles que aprenderam algo há muito tempo esquecem-se de como era difícil e por quais equívocos passaram.

Outra ótima maneira de começar a ministrar palestras é falar em meetups locais. Veja se algum está hospedando eventos com palestras-relâmpago, que são uma série de palestras curtas (frequentemente de cinco minutos). Esses eventos têm muito menos pressão para serem preparados, uma vez que terão entre cinco e uma dúzia de palestrantes em uma noite. Os eventos geralmente são bem receptivos aos palestrantes que falam pela primeira vez. Se não houver uma noite com palestras-relâmpago planejadas, mas você é um participante assíduo de um meetup local, recomende um evento desses aos organizadores!

**Gabriela de Queiroz: o início do R-ladies**

Quando me mudei para São Francisco em 2012, vindo do Brasil, fiquei impressionada com o número de recursos que encontrei. Rapidamente descobri os meetups e, por alguns meses, ia aos meetups todas as noites. Aprender e comer gratuitamente era uma combinação perfeita, especialmente para uma estudante sem muito dinheiro. Mas a maioria dos meetups não tinha um público diversificado. Não via ninguém como eu e não conseguia me sentir bem-vinda, por isso acabava em um canto, sem interagir muito.

Depois de algum tempo, decidi que era hora de retribuir à comunidade e começar meu próprio meetup. Eu era apaixonada pela linguagem R, mas não queria criar mais um grupo qualquer de R; queria um grupo onde eu (e os participantes) pudesse me sentir segura e bem-vinda, sem julgamento e que pudesse me enxergar no público. Assim nasceu o R-Ladies. Em outubro de 2012, organizei o primeiro evento, uma introdução à linguagem R (<http://bit.ly/rladies-first>), e apenas oito pessoas compareceram. Fiquei um pouco desapontada, mas contente por criar esse espaço e por ter sido corajosa o suficiente para ensinar uma linguagem de programação em uma língua estrangeira.

Durante quatro anos fui a única pessoa por trás do R-Ladies. Estava organizando, ensinando, anunciando e executando o site, além de procurar lugares e patrocinadores. Eu ia a conferências e eventos e falava sobre o grupo. Também era ativa nas redes sociais, tentando fazer tantas conexões quanto conseguisse. Infelizmente, a maioria dos meus empregadores não patrocinava meu trabalho; por isso, o R-ladies foi meu projeto paralelo, o que significa que passava noites e fins de semana trabalhando nele.

Ao liderar o R-Ladies, tive a oportunidade de conhecer muitas pessoas, algumas das quais nunca teria sonhado em encontrar na vida real. E como tive de ensinar nos eventos, fiquei mais confortável em falar na frente das pessoas.

Para quem deseja iniciar suas próprias comunidades, sugeriria o seguinte:



- *Defina um propósito e crie uma missão.* Qual é o objetivo desta comunidade? Aonde está tentando chegar? Por que está criando essa comunidade? Qual é a missão dela? Quem será o público? Pensar sobre essas perguntas ajudará seus futuros membros a compreender a razão por que devem se importar e por que devem participar. Também ajuda a influenciar decisões, como se quer se concentrar em um subgrupo específico, como o R-Ladies fez com mulheres e minorias de gênero, ou se quer atingir todos os interessados no tema.
- *Faça redes sociais, um site e um email.* Configure uma conta no Twitter, uma página no Facebook, um grupo no LinkedIn, um perfil no Instagram e qualquer outra rede social que tenha uma grande base de usuários. Também faça um site e um email para que as pessoas possam facilmente entrar em contato com você e saber mais sobre o grupo.
- *Crie um logo.* Ter um logo sensibiliza sua marca e, portanto, sua comunidade. Algumas pessoas têm uma memória visual melhor e recordarão do seu logo. Com um logo, é possível criar adesivos para notebooks, por exemplo. Os adesivos para notebooks são uma forma de se expressar, e também de expressar suas crenças e as comunidades das quais faz parte. É um grande sucesso!
- *Pense sobre o formato.* Serão principalmente palestras ou workshops? Será tudo presencial ou será uma comunidade online com eventos transmitidos ao vivo ou conversas em cafés? Se sua comunidade é de tecnologia, em que você queira empoderar seu público, um workshop seria um ótimo formato. A aprendizagem ativa é a melhor maneira de se aprender algo.
- *Utilize uma plataforma* (meetup.com ou eventbrite.com, por exemplo). É melhor facilitar a forma como as pessoas encontram e se inscrevem para seus eventos. Um site centralizado, como [www.meetup.com](http://www.meetup.com) ou [www.eventbrite.com](http://www.eventbrite.com), permitirá algum tráfego orgânico quando as pessoas estão pesquisando o tópico e ajuda a ficar a par do público esperado.

Construir uma comunidade requer tempo e esforço. Você provavelmente precisará trabalhar horas extras e nos fins de semana; então, é melhor ser algo pelo qual é apaixonado com uma missão na qual acredita. Apesar do trabalho, vale a pena! Ouvir as histórias de

sucesso, ver como sua comunidade mudou as comunidades locais em todo o mundo, especialmente em lugares carentes, é muito gratificante e uma fonte de grande alegria. Você sente que está fazendo algo para mudar o mundo para melhor. Boa sorte na sua jornada!

Por fim, você pode fazer com que pessoas o encontrem por meio do seu blog. Para as conferências que convidam palestrantes, se um dos organizadores lê um post de blog que se encaixa perfeitamente com o tema da conferência, eles podem chegar até você para ver se poderia ministrar uma palestra sobre o mesmo tema. Mesmo que não aconteça, os posts do blog são ótimas maneiras de mostrar aos organizadores da conferência que você é eficaz em se comunicar, mesmo se não tiver ministrado palestras antes.

Tal como os primeiros trabalhos de ciência de dados, as primeiras palestras são as mais difíceis de conseguir. Depois disso, muitas vezes sentirá um efeito de bola de neve, especialmente se a palestra foi gravada. Uma gravação é excelente porque as pessoas podem ver sua palestra e encontrá-lo, mas também porque algumas chamadas de apresentação pedem a gravação de uma palestra anterior.

### **14.3.2 Preparação**

Quando tiver um compromisso de palestra, você passará muito tempo preparando sua palestra. Se não tiver ministrado uma palestra pública antes, é fácil subestimar o tempo que leva. Sim, daria para ministrar uma palestra de última hora, fazendo cada slide com cinco pontos que são apenas seus pensamentos sobre o assunto e levar isso no dia do evento, mas é desrespeitoso com seu público e não demonstra seu melhor trabalho. Também não é o caminho para construir uma carreira bem-sucedida de palestrante.

É melhor praticar a palestra com uma pessoa de verdade, não apenas ler os slides para si mesmo. Encontre alguém em cuja crítica você confia e palestre para essa pessoa. A menos que esteja ministrando uma palestra muito técnica, provavelmente você não se importa se seu revisor tem alguma formação no assunto. Ele pode sugerir coisas gerais que farão sua palestra melhor, como enrolar menos ou não gesticular tanto.

Em geral, receberá um tempo-limite para sua palestra, mas este prazo pode ser complicado com base na sessão de perguntas e respostas posteriormente. Para calcular quanto tempo precisa para se preparar, reserve cinco minutos para perguntas e trabalhe com o tempo antes disso, mas é uma boa ideia se cronometrar ao palestrar. Tenha cuidado, pois existe a tentação de falar rapidamente diante das pessoas. Também é possível adicionar alguns slides extras no fim, caso tenha um tempo extra, e achar que se apressou um pouco demais na palestra principal. Mas, se você estiver com alguns minutos a menos, isso em geral é bom; haverá apenas uma pausa um pouco mais longa antes da palestra seguinte. O pior resultado é sua palestra ser longa demais e você ser cortado antes de terminar ou estourar o tempo, atrapalhando o próximo palestrante.

Por causa de todo o trabalho que dá planejar uma palestra, recomendamos fortemente reutilizar palestras. É muito improvável que o público seja igual, especialmente se as palestras forem em cidades diferentes ou ministradas em um evento com vários eixos (onde os participantes podem se encaminhar para outra palestra que esteja acontecendo ao mesmo tempo). Embora seja lisonjeiro pensar que todos assistiram à gravação da sua palestra, a maioria das pessoas não fará isso.

No dia da palestra, reúna seus apoiadores. Esse grupo não precisa ser limitado a seus amigos e colegas da ciência de dados; convide familiares, cônjuges e o amigo do seu prédio. Se for um evento pago, veja se os organizadores oferecem um passe a um familiar ou cônjuge para assistir à palestra. O avô de Emily foi a várias de suas palestras gratuitamente (para a alegria do público). É bom saber que pelo menos parte do público definitivamente está torcendo por você.

## **14.4 Como contribuir para o código aberto**

Para aqueles que apreciam a ideia de fazer parte de uma comunidade, mas não gostam da ideia de estar em uma sala com outras pessoas nessa comunidade, o código aberto pode atender a essa necessidade. Contribuir para o código aberto permite compartilhar ideias e desenvolve um sentido de comunidade entre as pessoas com a mesma paixão. Criar um projeto em

código aberto pode gerar muito interesse, já que as pessoas avançarão em novas direções que possivelmente não tenha considerado. Da mesma forma, você pode expandir o trabalho de outra pessoa para gerar um projeto totalmente novo.

As linguagens R e Python prosperam porque voluntários estão sempre as expandindo e refinando. Nas seções a seguir, discutimos como se tornar um desses voluntários; você também pode contribuir financeiramente para as organizações que patrocinam alguns dos principais desenvolvimentos. Embora R e Python possam não ter custos para serem utilizados, elas têm custos de manutenção e desenvolvimento. A R Foundation, a Python Software Foundation e a NumFOCUS são três organizações (as duas últimas são instituições de caridade registradas nos EUA) às quais se pode fazer doações para apoiar o desenvolvimento contínuo das linguagens.

#### **14.4.1 Como contribuir para o trabalho de outras pessoas**

Entrar em um projeto de código aberto pode dar a sensação de estar mexendo no armário de outra pessoa. É o espaço delas, e você se sente um intruso, mas o código aberto foi construído exatamente para esse propósito, e você tem de passar por esse sentimento. Em vez disso, imagine que os projetos de código aberto são como organizar um jantar gigante. Provavelmente não irá querer ser responsável pelo prato principal ainda, mas muitos trabalhos precisam ser feitos; você pode ajudar a colocar a mesa, garantir que todos tenham água ou guardar os pratos depois. Se for respeitoso e entusiasmado, a maioria dos criadores e mantenedores receberá sua ajuda.

Um ótimo lugar para começar a contribuir é com a documentação. Veja como a documentação é fornecida para um pacote de que você gosta. Talvez veja algo incompleto, que não está claro ou que induz ao erro. Mesmo para a correção de um erro de digitação vale a pena fazer uma solicitação de pull no GitHub. Os criadores de pacotes e bibliotecas adoram ter mais trabalhos escritos sobre eles. Esse trabalho economiza tempo e, como alguém que aprendeu recentemente a usar essas ferramentas, você terá uma perspectiva melhor sobre o que motivará e ensinará novos usuários.

Se quiser contribuir com código, não comece a logo reescrever coisas ou enviar uma nova função. Se o projeto for grande, pode existir um guia sobre como contribuir ou um código de conduta. Se não for o caso, observe o repositório por um tempo para entender o fluxo. Observar o repositório também dirá se o projeto é mantido ativamente ou se fica inativo por longos períodos. Se você decidir que deseja começar a contribuir com o código, comece compartilhando o que gostaria de adicionar ou alterar. Assim, você pode obter comentários dos mantenedores antes de realizar muito trabalho.

Trabalhar em código aberto é uma das melhores maneiras de desenvolver suas competências técnicas, especialmente se nada no trabalho exigir que coopere com um grande grupo de pessoas. Talvez em seus repositórios GitHub de trabalho, você não use ramificações (branches), mensagens informativas de commit ou de marcação (tagging). Tudo bem, mas, ao entrar em um projeto com centenas de problemas e dezenas de pessoas trabalhando ao mesmo tempo nele, esse trabalho extra começa a fazer mais sentido. Esses tipos de práticas adicionam restrições extras, quer você esteja trabalhando dentro de um guia de estilo ou os mantenedores não adicionem um recurso que você criou, pois ele não tem desempenho suficiente. Eles serão os tomadores de decisão finais até que você crie um projeto próprio. Embora possa ser frustrante, aprenderá muitas práticas recomendadas que podem ser aplicadas no seu próprio trabalho.

**Reshama Shaikh: hackathons**

Contribuir para o código aberto pode parecer enigmático e assustador. Os sprints de código aberto, às vezes chamados de *hackathons*, são eventos estruturados que fornecem um espaço acolhedor para iniciantes. Os sprints são tipicamente eventos de um ou dois dias em que os participantes trabalham em problemas em aberto enviados ao repositório GitHub de uma biblioteca Python ou R. Esses problemas podem estar relacionados à documentação, correções de bugs, testes, solicitações de recursos e muito mais.

Os benefícios de participar de sprints de código aberto são muitos:

- A maioria dos colaboradores de código aberto é voluntária; portanto, o envolvimento da comunidade é essencial e bem-vindo.
- É um evento ativo e prático que desenvolve competências de engenharia e programação.
- Contribuir para o código aberto é uma excelente oportunidade de aprendizagem que faz suas competências em ciências de dados avançarem e constrói seu portfólio.
- Oferece uma oportunidade valiosa de formar uma rede de contato com outros cientistas de dados e contribuidores experientes.

Um sprint bem organizado utilizará o tempo das pessoas de forma eficiente. A preparação assegura que os colaboradores iniciantes são capazes de deixar o sprint tendo realizado algo. Procure um repositório central disponível de recursos e trabalho de preparação, que inclui documentação contributiva, instruções de instalação de R ou Python, ferramentas para se inscrever antes do evento (como uma conta GitHub ou plataforma de mensagens) e uma lista de problemas em aberto, especialmente preparada para os participantes do sprint. Tenha em mente que as pessoas que organizam esses sprints são voluntárias; se achar que alguns desses itens estão faltando, ofereça ajuda. Organizar sprints de código aberto também contribui para o código aberto.

O objetivo de um sprint de código aberto é enviar solicitações de pull (PRs – Pull Requests) que resolvem problemas em aberto. Submeter uma PR é um processo de idas e vindas, sendo comum levar diversas semanas para ser mesclada. Reserve um tempo pós-sprint (normalmente, 5 a 10 horas) para acompanhar o trabalho e ver uma PR chegar ao estado de mesclado, que é representado no repo do GitHub por um lindo ícone roxo.

Se estiver interessado em organizar um sprint você mesmo, escrevi um guia detalhado em meu blog em <https://reshamas.github.io/how-to-organize-a-scikit-learn-sprint> (em inglês).

### **14.4.2 Como fazer seu próprio pacote ou biblioteca**

Quando você se encontrar copiando funções entre projetos ou as enviando por mensagem para seus colegas de trabalho, pode ser a hora de criar um pacote ou uma biblioteca. Um pacote permite armazenar funções em um único lugar, compartilhá-las facilmente e aplicar práticas recomendadas, como testar o código. Muitas empresas têm pacotes internos com funções para fazer com que a cor de seus desenhos seja a mesma da empresa, acessar dados ou resolver problemas comuns. Se achar que outros podem estar enfrentando o mesmo problema, pode compartilhar seu pacote no GitHub para que outras pessoas possam baixá-lo e usá-lo.

Antes de tentar fazer que o público use algo, você precisa garantir que todo seu código está em ordem. Só porque algo funcionou bem para você ao executar uma tarefa não significa que funcionará sob o estresse da utilização pública. Se seu código é algo que acabou de ser feito, mas você não tem certeza de como funciona, não convide as pessoas para utilizá-lo ainda. Tornar seu pacote mais amplamente útil pode exigir programação mais avançada à medida que refina ou adapta o pacote para caber em um caso generalizado. Certifique-se de que seu trabalho subjacente foi lido por alguém em quem confia. Os usuários não olharão debaixo da capa; por isso, se disser que é uma Ferrari, ficarão chateados quando se tratar de um carrinho de golfe por quase todo o tempo.

Quando tiver testado e revisado seu código, ainda é preciso trabalhar para que as pessoas possam descobri-lo. Você pode divulgá-lo nas redes sociais ou em seu blog, mas, mesmo assim, isso pode ser um processo lento. Não espere virar uma estrela da noite para o dia; é melhor trabalhar no início com menos usuários do que ter sucesso imediato e perceber que cometeu um erro em seu código subjacente. Pode levar algum tempo para que algo seja adotado, se de fato for, mas até mesmo tentar divulgar um bom trabalho é uma boa ação. A recompensa do sucesso é também uma maldição, claro: se as pessoas começarem a confiar no seu projeto, torna-se muito difícil



parar de desenvolvê-lo. Você receberá relatórios de bugs e solicitações de recursos e terá de considerar seriamente se fará uma mudança que quebrará relatórios que usaram a versão antiga de uma função.

**TOXICIDADE NO CÓDIGO ABERTO** Comunidades de código aberto podem ser tóxicas. As pessoas tiveram experiências negativas nas quais são discriminadas, assediadas, depreciadas ou simplesmente não se sentem bem-vindas devido à raça, ao gênero, à etnia ou à sexualidade. Felizmente, muitas comunidades estão reconhecendo esse fato e trabalhando ativamente para tornar o ambiente mais inclusivo. Guido van Rossum, criador do Python, comprometeu-se a fazer a mentoria somente de mulheres e minorias (<http://mng.bz/9wPo>). Alguns criadores de projetos etiquetam problemas “amigáveis para principiantes” ou “primeira vez” para encorajar aqueles que são novos a contribuir com o código aberto. Embora sempre deva priorizar sua saúde mental e emocional, muitas pessoas, incluindo minorias, tiveram somente experiências positivas no código aberto; uma experiência ruim não é inevitável.

## 14.5 Como reconhecer e evitar a exaustão

Não somos especialistas em saúde, por isso recorremos à definição de *exaustão (burnout)* da Organização Mundial de Saúde: uma “síndrome resultante do estresse crônico no local de trabalho que não foi gerida com sucesso”. Lista três sintomas como “sentimentos de exaustão ou esgotamento de energia”, “maior distância mental do trabalho ou sentimentos de negativismo ou cinismo relacionados ao trabalho” e “redução da produtividade profissional” ([https://www.who.int/mental\\_health/evidence/burn-out/en](https://www.who.int/mental_health/evidence/burn-out/en)). Por agora, vamos nos concentrar no estresse que não provém do seu trabalho de tempo integral, mas, sim, do trabalho extra relacionado à carreira com aquilo que faz em paralelo.

Escrever este livro era algo que fizemos completamente à parte de nossos trabalhos de tempo integral (e, para Jacqueline, a parte de criar uma criança). Certamente, às vezes temos ciúme quando colegas vão para casa e não fazem nada relacionado à ciência de dados. Para nós, isso nos ajuda a voltar ao motivo pelo qual decidimos assumir este trabalho extra e ver se ainda estamos trabalhando na direção de nossos objetivos. Com este livro, nunca se tratou de ganhar dinheiro. (Passar as horas de escrita com consultoria teria sido muito mais lucrativo.) Em vez disso, queríamos escrever este livro para ajudar os aspirantes e cientistas de dados iniciantes, e essa missão nos manteve motivadas. Foi especialmente útil ver o impacto

positivo ao longo do caminho à medida que lançamos os capítulos.

Caso sinta-se exausto, comece se perguntando se há maneiras de diminuir o ritmo. Algo que é útil de lembrar é que, depois de criar algo, não é necessário permanecer ativo. Se mantiver um blog, talvez queira escrever um novo post ocasionalmente, mas não precisa escrever com a mesma frequência que fazia no início. É mais provável que alguém que esteja visitando diga: “Uau, esses seis posts me ajudaram muito” do que: “Puxa, agora ela só posta a cada seis meses”.

Na cultura atual de correria e no louvor de manter-se muito ocupado, pode parecer que qualquer momento que não se é produtivo é tempo perdido. Isso é muito prejudicial! Todos precisamos de tempo para reiniciarmos. Coisas como ir para a academia e sair com amigos são boas para esse fim, mas também pode ser tempo assistindo TV ou simplesmente relaxando. Continue criando espaço para os passatempos que não têm nada a ver com a ciência de dados ou com ganhar dinheiro para que não sinta como se toda sua vida girasse em torno do trabalho.

Você pode adicionar muito estresse tentando acompanhar o ritmo de outras pessoas. É comum dizer que as redes sociais são o ponto alto de outra pessoa, e não se deve compará-las com sua vida inteira. Da mesma forma, só porque alguém é um criador produtivo de pacotes ou posts de blog não significa que você precise acompanhar essa pessoa. Para algumas, criar pacotes, ministrar palestras ou escrever posts de blog é parte ou até mesmo uma obrigação de trabalho! O melhor a se fazer para a carreira é trabalhar de forma sustentável o suficiente para se manter no jogo a longo prazo.

## **14.6 Entrevista com Renee Teate, diretora de ciência de dados na HelioCampus**

Renee Teate é conhecida no Twitter pelos seus mais de 50 mil seguidores como Data Science Renee em @becomingdatasci. Ela também criou um podcast, blog ([www.becomingdatascientist.com](http://www.becomingdatascientist.com)) e [www.datasciguide.com](http://www.datasciguide.com) um diretório online de aprendizado em ciência de dados. Ela palestra regularmente e organiza conferências.

## **Quais são os principais benefícios de se estar nas redes sociais?**

O Twitter beneficiou-me de muitas formas. Todos os convidados em meu podcast são pessoas que realmente conheci pelo Twitter. Encontrei pessoas que achei que estavam tuitando sobre coisas interessantes e pensei que, se tuitavam coisas interessantes, também poderiam falar sobre coisas interessantes. Fiz uma lista e enviei várias mensagens diretas de uma só vez, pensando que talvez metade estaria interessada e eu poderia agendar algo no futuro. Bem, cada um deles disse sim!

Regularmente pedem-me para palestrar em conferências e meetups pelo Twitter. Quando libero conteúdo, sei que há um público para ele, e ele irá se engajar. Conheci tantas pessoas interessantes! Além de formar essa rede de contato, também o utilizo para aprender. Na verdade, escrevi um post de blog no início sobre como uso o Twitter para aprender coisas, e ele está focado principalmente em aprender o linguajar da área. Se começar a seguir pessoas em uma determinada área e ler os artigos que estão colocando no link, aprende-se toda essa terminologia. Se alguém tiver dito algo que não sabia, eu simplesmente pesquisava e descobria o que era. Muitas vezes, havia um link para um tutorial ou artigo sobre o tópico, então, realmente, me ajudou com a aprendizagem também.

## **O que diria às pessoas que falam que não têm tempo para se envolver com a comunidade?**

Compreendo totalmente, em especial pessoas que têm responsabilidades fora do trabalho, como cuidar de uma criança ou de outro parente. Quando estava no programa de mestrado e trabalhando em tempo integral, não fazia mais nada. Nesses casos, aconselharia a encontrar uma comunidade online com a qual interagir de forma assíncrona. Sempre que tiver um pouco de tempo, como se estivesse em uma sala de espera, leia e responda a alguns tweets ou marque artigos interessantes para ler mais tarde. Mesmo que só possa realizar um evento por ano, escolha uma conferência relacionada à ciência de dados ou com sua área específica e faça um esforço para comparecer. Posteriormente, dá para acompanhar as pessoas que conhecer no evento pelo LinkedIn ou por outras redes sociais e, às vezes, pode se tornar um pequeno grupo para se apoiarem no processo de aprendizagem,

além de compartilhar recursos.

## **Há valor em produzir apenas uma pequena quantidade de conteúdo?**

Com certeza! Penso que mesmo se escrever apenas um post de blog ajudará a solidificar o tópico em sua mente porque você aprende ajudando outras pessoas a compreenderem algo. Referencio alguns dos meus posts antigos do blog há anos. Quando escrevo no blog, tento fazer algo aplicável a um nível muito geral e benéfico para os alunos, para que possam voltar a consultá-lo várias vezes sem que rapidamente fique desatualizado. Para o meu podcast, só gravei dois episódios no último ano e meio. Tornei-me bem ocupada e tinha começado meu trabalho na HelioCampus e deixei o podcast de lado, mas foi realmente difícil voltar a ele depois de me comprometer com outras coisas. Ele ainda existe, e pretendo fazer mais episódios. Mas parei de me sentir culpada sobre fazer um grande intervalo. Percebi que os episódios que estão lá ainda são úteis para as pessoas, e sempre posso ouvi-los mais tarde.

## **Você ficou preocupada na primeira vez que publicou um post no blog ou palestrou?**

Sim, fiquei ansiosa em fazer algo assim pela primeira vez, pois, quando as pessoas pesquisam seu nome, elas vão encontrá-lo e associá-lo com você. É claro que é um pouco angustiante. Algo que me dei conta é que o tipo de post de blog que já existia e que recebi um retorno não era sempre o mais tecnicamente avançado ou perfeitamente escrito. Eu costumava ler blogs que descreviam algo que eu queria aprender de uma maneira ligeiramente diferente da que tinha ouvido antes e, de repente, o material fazia sentido. Há sempre alguém que se beneficia daquilo que você publica.

Também aprendi a não me importar muito com as pessoas do contra. Eles sempre existirão. Vi fazerem comentários negativos a pessoas que fazem ciência de dados há anos. Existem muitas formas de abordar uma análise e talvez uma delas seja melhor do que outra por algumas razões, mas não significa que seu caminho não seja uma boa maneira. Às vezes, basta parar de ouvir os críticos.

## Resumo

- Há quatro maneiras que recomendamos para se envolver com a comunidade da ciência de dados: fazer um blog e construir seu portfólio de ciência de dados, participar de conferências, palestrar e contribuir ao código aberto.
- Lembre-se de que não é preciso fazer nenhuma atividade na comunidade para ter uma carreira bem-sucedida; escolha o que funciona para você e não se preocupe em acompanhar outras pessoas.

## CAPÍTULO 15

# Como sair bem do seu emprego

Este capítulo abrange:

- Decidir quando sair de um emprego
- Compreender como a busca de emprego difere do seu primeiro emprego de ciência de dados
- Apresentar o aviso-prévio e gerenciar a transição

Já se foi o tempo em que alguém passava 40 anos em uma mesma empresa e, ao se aposentar, recebia um relógio de ouro e a aposentadoria. Na maioria das áreas, é comum agora mudar de empresa pelo menos algumas vezes na carreira e, na tecnologia, as pessoas podem trocar de emprego a cada dois anos. Há muitas boas razões para sair de um emprego: poderia estar buscando um aumento de salário, responsabilidades diferentes, aprendizagem acelerada ou simplesmente algo novo. Decidir se está interessado em um novo emprego é o primeiro passo, mas há obstáculos mentais adicionais a superar entre essa etapa e realmente fazer algo com relação a ela.

Há sempre a incerteza de sair de uma função que você conhece para uma nova. Não importa quanta pesquisa realize ou quantas perguntas faça na entrevista, nunca é possível saber como será de verdade até começar o trabalho. Você pode entender as coisas maiores – o salário, o tamanho da empresa e a estrutura da equipe de dados –, mas não saberá como irá se sentir no dia a dia até vivenciá-lo. Além disso, é provável que seu trabalho atual não seja de todo horrível. (Se for, recomendamos voltar ao Capítulo 9, onde discutimos o que fazer se o trabalho for terrível ou se o ambiente for tóxico.) Você provavelmente gosta de alguns colegas de trabalho, sabe onde conseguir ajuda e está confortável navegando pelos dados. Também pode pensar em algumas coisas que podem ser melhores. Mas qual é a garantia

de que um novo trabalho que você acha que seria melhor será mesmo melhor ou que não será algo ainda pior? Vale mesmo a pena o risco e o tempo para encontrar um novo emprego?

Essas dúvidas incômodas podem retardar seu progresso na busca por um emprego mesmo depois de decidir que quer sair. Provavelmente também está lidando com muita incerteza. Como abordar sua segunda busca de emprego na ciência de dados? Se receber uma proposta que quer aceitar, como você comunica ao seu gerente? Melhor será reunir-se em particular com cada colega com quem trabalhou em um projeto para dizer-lhes que está se demitindo? Se receber uma contraproposta, deve aceitá-la? O que fazer nas últimas semanas depois de dar o aviso-prévio? Sua procura por um segundo (ou terceiro, ou quarto) emprego não precisa ser algo que o amedronte, basta saber que procurar por algo melhor não o impede também de mudar de ideia e preferir manter seu emprego atual.

Levantamos muitas perguntas que fazíamos quando considerávamos a busca por novos empregos. O enorme volume de incerteza pode paralisar todos, exceto os candidatos mais focados, mas não tema: estamos aqui para transformá-lo em um deles.

Neste capítulo, dividimos a situação de sair bem do emprego em três partes: decidir sair, iniciar a busca de um novo emprego e apresentar o aviso-prévio. Algumas dessas orientações aplicam-se a qualquer emprego, mas também discutimos alguns aspectos que são mais exclusivos da ciência de dados. A mudança de empregos na ciência de dados é uma experiência comum e geralmente recompensadora; muitas pessoas mudam de emprego a cada um a três anos, o que lhes permite experimentar novas áreas da ciência de dados e aumentar significativamente seus salários e outros benefícios. Este capítulo irá ajudá-lo a fazer essa transição da maneira mais fácil e o menor estresse possível.

## **15.1 Decidir sair**

Infelizmente, na maioria das vezes, ninguém sabe com 100% de certeza quando é o melhor momento de sair do emprego. Não há uma bola 8 mágica que lhe diga o que fazer ou até mesmo um conjunto de perguntas

que possa responder e que o ajuda a tomar uma decisão definitiva. No Capítulo 8, discutimos como escolher entre duas boas opções de vida ao decidir-se entre propostas, e esse mesmo estilo de raciocínio aplica-se aqui. No fim do dia, só dá para você fazer o melhor com a informação que tem em mãos, e muito poucas decisões são completamente irreversíveis. Também é sempre possível você sair logo do novo emprego, pois não está assinando um contrato de 100 anos.

### **15.1.1 Faça o balanço do seu progresso de aprendizagem**

Em que momento saber que está na hora de buscar um novo emprego? Nosso maior conselho é garantir que esteja sempre aprendendo. Infelizmente, pode ser comum que sua aprendizagem fique mais lenta ao permanecer em uma mesma função. Nos primeiros meses, você aprende de tudo. É praticamente impossível *não* aprender algo; no mínimo, aprenderá sobre os dados da empresa, novas competências técnicas de colegas e a trabalhar com stakeholders. Mas se continuar fazendo a mesma coisa depois de um ano ou dois, talvez fique estagnado.

Ao ficar mais confortável com os aspectos do dia a dia do trabalho, veja o que é possível fazer para encontrar maneiras de melhorar suas competências não técnicas. Veja se pode assumir uma equipe (ou pelo menos um estagiário) e trabalhar em sua capacidade de gerenciamento. Embora o trabalho que a empresa precisa que você desenvolva possa ser limitante, ao ficar experiente nele, normalmente há como encontrar mais tempo para ampliar seu portfólio. Talvez possa colaborar com a equipe de engenharia de dados para aprender a construir um pipeline por conta própria em vez de depender apenas dos engenheiros. No entanto, tomar a iniciativa assim não é para todos; por vezes, as pessoas precisam de motivação externa para enfrentarem novos desafios. Caso você se encontre estagnado, é um sinal de que essa pode ser a hora de uma mudança de cenário.

Algo animador sobre a ciência de dados é que há sempre mais a aprender, mas esse fato também torna o trabalho desafiador. Se você não crescer, encontrar seu próximo emprego será mais difícil. Espera-se também que um cientista de dados sênior tenha competências visivelmente diferentes das de um cientista de dados júnior, tanto em termos de amplitude quanto de



profundidade. Ao longo deste livro, salientamos que não é preciso nem possível saber tudo sobre ciência de dados, todavia, espera-se que saiba mais à medida que adquire experiência.

### **15.1.2 Verifique como você se alinha com seu gerente**

Antes de cortar laços e sair correndo, garanta ter feito o possível para dizer ao seu gerente o que gostaria de mudar. O que parece ser um problema não solucionável pode, na verdade, ter uma solução. Talvez esteja atolado em tarefas difíceis que não podem ser automatizadas, mas que já não são mais desafiadoras. Seu gerente pode lhe dizer que pode contratar um estagiário para fazer esse trabalho. Esse estagiário obtém uma experiência de aprendizagem, e você consegue trabalhar um pouco fora dessa estagnação e ganhar experiência de mentoria. Ou, talvez, a equipe de ciência de dados faça a maior parte das análises, mas você quer muito começar a colocar machine learning em produção. Seu gerente pode fazer um “bootcamp” com uma equipe de engenharia durante alguns meses; você pode aprender alguns princípios básicos da engenharia, contribuindo simultaneamente com seus conhecimentos analíticos.

Outra pergunta a se fazer é saber como seus objetivos estão alinhados com os do seu gerente. Philip Guo, professor assistente de ciências cognitivas da Universidade da Califórnia – San Diego, escreveu um post de blog chamado “Whose critical path are you on?” (“Você se encontra no caminho crítico de quem?”) (<http://www.pgbovine.net/critical-path.htm>), no qual discute a importância de conhecer o caminho crítico do seu chefe (ou mentor) e se ele se alinha com o seu. *Caminho crítico* aqui significa “o caminho do trabalho que é crítico para o avanço ou a realização da carreira em um determinado momento”. É sobre o sucesso do seu chefe estar atrelado ao seu. Os gerentes têm tempo e energia limitados e, se seus caminhos críticos se sobrepuserem, eles terão maior probabilidade de se concentrarem em você.

Saber o quão bem você está alinhado com os objetivos do seu gerente requer saber quais são seus próprios objetivos de carreira. Não estamos falando de um plano de dez ou mesmo de cinco anos; em uma área tão nova e em rápido desenvolvimento, é impossível saber quais as oportunidades

que estarão disponíveis no futuro. Mas como você quer passar os próximos anos? Esperamos que tenha pensado bastante nessa questão durante sua primeira busca de emprego, mas talvez as coisas tenham mudado... Você pode ter desejado estar em uma grande equipe de ciência de dados, mas, agora que está em uma há alguns anos, gostaria de trabalhar em diferentes tipos de projetos em vez de ficar preso a um. Ou poderia focar em sua família e buscar um trabalho que lhe permita trabalhar das 9 às 17h do que um emprego em uma startup que demande muitas horas.

Em resumo, alguns fatores importantes a considerar ao pensar em encontrar um novo emprego são:

- Você está aprendendo em sua função atual?
- Tentou melhorar sua experiência diária discutindo seus problemas com seu gerente?
- Seu gerente está focado nas suas necessidades e no avanço da sua carreira?
- Passou tempo pensando sobre o que você faz atualmente e não quer fazer em seu próximo emprego?

**Sair do seu emprego sem ter outro em vista**

Talvez queira uma folga significativa entre sair e começar em outro emprego. A maioria dos novos empregadores quer que você comece assim que possível; embora geralmente consiga uma ou duas semanas entre sair de seu emprego e começar no novo (em especial se já tem férias planejadas), é improvável conseguir mais do que isso. Se você estiver sonhando com um mochilão de três meses pela Ásia, provavelmente terá de sair do emprego sem qualquer outro em vista.

Sair da empresa sem outro trabalho é arriscado. Há o risco financeiro de talvez não ter economias suficientes para um período indeterminado sem renda. Está disposto a (ou é capaz de) contar com empréstimos a curto prazo de familiares ou dinheiro da renda do cônjuge? O outro ponto é que é mais fácil encontrar um emprego quando se está empregado. Uma razão para isso é o preconceito injusto entre os gerentes de contratação contra aqueles que estão desempregados. Outra é que sua posição para negociar fica mais fraca: o que o novo empregador propuser será mais do aquilo que está ganhado; então, é mais difícil pedir uma remuneração mais alta. Mas a outra razão é que, se tiver tirado meses de folga, talvez você não tenha mantido suas competências e possa estar enferrujado para entrevistas técnicas.

Se quiser tirar uma folga, é extremamente útil ter uma forte rede de contatos na ciência de dados: pessoas que estão familiarizadas com seu trabalho e que podem recomendá-lo a um gerente. Também se prepare por algum tempo antes de ser entrevistado para polir suas competências técnicas. No geral, a menos que esteja em um ambiente de trabalho tóxico, recomendamos não sair de um emprego sem ter outro, a menos que haja um plano com o qual esteja animado para o tempo de espera possivelmente longo entre um emprego e outro.

## **15.2 Como a busca de emprego difere do seu primeiro emprego**

Muitos dos princípios básicos para encontrar o seu segundo emprego na ciência de dados são os mesmos que os do primeiro. Mas há algumas vantagens significativas agora que você tem a experiência de trabalhar em ciência de dados:

- Haverá mais recrutadores que chegam em você. Para aumentar o interesse dos recrutadores, você pode configurar no seu perfil do LinkedIn para mostrar aos recrutadores que está aberto a oportunidades de emprego (e não se preocupe, o LinkedIn toma medidas para garantir que seus empregadores não vejam isso).
- Você aprendeu mais sobre os aspectos de seu trabalho de que gosta e aqueles dos quais não gosta. É cedo o suficiente na sua carreira para mudar sua especialidade de direção: se tiver feito muito trabalho de engenharia de dados, mas não gostou, pode ir a uma empresa maior que tenha engenheiros de dados para fazer esse trabalho.
- Será mais fácil chegar à primeira etapa de seleção. Muitos empregadores usam como uma ferramenta de seleção rápida se alguém já teve o mesmo cargo (ou algo muito similar antes).
- Em teoria, sua rede de contatos na ciência de dados está mais desenvolvida. (Caso contrário, volte ao Capítulo 14.)
- Se você ainda estiver empregado, é melhor não postar no LinkedIn ou no Twitter que está buscando um novo emprego, mas dá para começar a

avisar aos poucos a algumas pessoas confiáveis para que saibam que está nessa busca. Talvez possam recomendá-lo nas empresas em que atuam ou colocá-lo em contato com alguém que eles saibam que está contratando.

Não tenha medo de se candidatar a vagas, mesmo que esteja razoavelmente satisfeito com seu trabalho no momento. Há muitas razões para que se convencer a não fazer uma mudança. Talvez pense não ter as competências que acha que os demais possuem (“E se eu tiver anos de experiência e não passar no estudo de caso técnico?”). É apenas a síndrome do impostor falando (e se tiver aprendido alguma coisa com este livro, então, que seja lutar contra a voz insistente dizendo que você não é tão bom quanto todos os outros). Se você não passar no teste técnico, não significa que seja um fracassado ou um cientista de dados “falso”. Muitas perguntas de entrevistas mal formuladas não julgam a capacidade com precisão. Além disso, a ciência de dados é tão ampla que talvez as perguntas estivessem em uma área na qual você não tenha trabalhado antes.

Também pode estar preocupado sobre as repercussões em sua vida social com seus amigos do trabalho ou a possibilidade de que um novo trabalho poderia levá-lo mais longe de casa. Mas, independentemente das suas preocupações, fará um desfavor a si mesmo se não estiver aberto a um processo de seleção que possa melhorar sua carreira.

### **15.2.1 Decidir o que você quer**

O primeiro passo na sua busca por emprego é fazer uma lista do que gostou no seu trabalho atual. *Designing Your Life: How to Build a Well-Lived, Joyful Life*, por Bill Burnett e Dave Evans (Knopf, 2016), sugere que, durante uma semana, você faça anotações antes e depois de cada atividade sobre quanto você pensou que iria gostar e quanto realmente gostou. Você odeia ter reuniões algumas horas por dia ou realmente gosta de reuniões porque trazem certa estrutura ao seu dia? Se estiver em uma equipe de ciência de dados distribuída, gostaria de se reportar a um gerente de ciência de dados? É possível usar essa lista para planejar sua busca. Encontre uma empresa que valorize as mesmas coisas que você ou que tenha a estrutura que está buscando. É melhor não se candidatar a uma empresa na qual

enfrentaria os mesmos problemas que estão lhe causando infelicidade agora.

Na sua busca, pode ocorrer o problema do título do cargo. No Capítulo 5, falamos sobre não se preocupar com títulos. Os cientistas de dados têm muitos títulos diferentes, incluindo analista de dados, cientista de pesquisa, engenheiro de machine learning e analista de produto. Analista de dados é o título mais comum e pode ser visto como uma função júnior. Se for um cientista de dados, estaria disposto a aceitar um título de analista de dados sênior? Se for um analista de dados, deveria se concentrar em avançar para o título de cientista de dados em sua nova função?

Aprender é ainda o fator o mais importante na sua busca. O que faria na sua nova função? Pense em termos de cinco anos, não nos próximos dois. O que irá prepará-lo melhor para o sucesso a longo prazo? É possível trabalhar como analista de dados sênior e, depois, fazer a transição para uma função de cientista de dados, por exemplo? Trabalhar em uma empresa de tecnologia menor permitirá que você aprenda a trabalhar com dados da web, preparando-se para ir a uma empresa de tecnologia de grande porte?

Quando estiver considerando suas opções, é preciso proteger seu valor de mercado. De forma justa ou não, o título de cientista de dados ainda é normalmente visto como mais prestigiado do que o de analista de dados, e a função de um analista de dados sênior pode pagar menos do que a de um cientista de dados. É melhor colocar na balança essas considerações ao pensar sobre sua próxima função.

### **15.2.2 Entrevistas**

Depois de começar a candidatar-se a empresas e a ser chamado para entrevistas, terá de responder à pergunta: “Por que razão quer sair do seu emprego?”. Se tiver conseguido seu emprego atual depois de sair da universidade ou de um bootcamp, você não foi questionado sobre isso.

Uma boa resposta é que está em busca de desafios. Outra boa estratégia é fingir que a pergunta é: “Por que você quer trabalhar conosco?”. Se imaginar isso, sua resposta será positiva (“Ouvi coisas incríveis sobre sua equipe de machine learning e quero muito aprender.”) em vez de negativa (“Meu chefe anterior insistiu em exibir nossos resultados de experiência em

gráficos de pizza.”). Se der uma resposta mais específica, certifique-se de que seja coerente com o novo empregador. Não diga algo como: “Estou buscando uma equipe com cientistas de dados sênior com quem possa trabalhar” se a empresa não tiver uma! Evite falar mal do seu trabalho atual a todo custo; alguns empregadores veem esse comportamento como desclassificatório, não importa quão mal tratado você seja em sua empresa atual.

Só porque está saindo da sua empresa não significa que você não deva se orgulhar do trabalho que desenvolveu lá. Fale sobre projetos nos quais trabalhou ou sobre competências que aprendeu. Provavelmente há acordos de confidencialidade, portanto não poderá mostrar seu código ou falar sobre os parâmetros do algoritmo de recomendação que construiu, mas aborde as contribuições que deu de uma maneira geral. Uma boa resposta não específica é: “Criei um chatbot em Python que gerou respostas para perguntas frequentes de clientes, diminuindo em cinco minutos o tempo médio que um representante de atendimento ao cliente precisava passar com cada cliente e aumentando a satisfação do cliente em 20%”. Por outro lado, se trabalhar em uma empresa privada, e dizer: “Realizei testes A/B que aumentaram a receita total da empresa de 20 milhões de dólares para 23 milhões de dólares”, é uma resposta inadequada, pois está revelando informações financeiras privadas.

Talvez esteja vendo tarefas que usam tecnologias diferentes, sejam elas de distintos provedores de nuvem, dialetos de SQL ou linguagens de programação principais. Nesse caso, é melhor usar as estratégias similares àquelas de quando enquadrou suas experiências de trabalho em termos de competências que são transferíveis à ciência de dados. Suponha que tenha trabalhado com linguagem R, mas a empresa usa Python. É possível dizer algo como: “Sei que levará um pouco de tempo para eu me adequar à sintaxe em Python, e já comecei a fazer um curso online. Mas em meus quatro anos de programação em R desenvolvi aplicativos web, construí pacotes e analisei grandes conjuntos de dados, os quais me farão um bom programador em Python rapidamente”.

Mencionamos a síndrome do impostor neste capítulo, mas você realmente precisa ter cuidado com isso ao se preparar para entrevistas. Quando estiver

buscando seu primeiro emprego saindo da faculdade ou quando está fazendo a transição para uma nova carreira na área da ciência de dados, é fácil dizer: “Ainda não aprendi isso”. (Pelo menos, é fácil de se convencer de que é fácil.) No entanto, quando você já está no meio do caminho, pode ter vergonha por não saber algo. Se não souber algo em uma entrevista, não tenha medo de admiti-lo. É possível dizer que não teve a oportunidade de usá-lo ou que espera aprender mais sobre o assunto, mas ainda não fez parte de seu trabalho. Suponha que seja questionado sobre algoritmos de machine learning, mas você trabalha com modelagem estatística, SQL, limpeza de dados e com stakeholders, porque os dados que estão em uma escala que machine learning é feito por engenheiros dedicados a isso. Ninguém sabe de tudo, e esperamos que você tenha feito um bom trabalho como cientista de dados até agora; tenha fé nisso. Dá para mostrar o trabalho que realizou e, se já estudou um tópico específico, mesmo que não tenha utilizado esse conhecimento recentemente, poderá avançar com mais rapidez. É sempre melhor demonstrar vontade de aprender do que tentar fingir informações sobre a sua trajetória.

## **15.3 Encontrar um novo emprego enquanto estiver empregado**

Se sua trajetória para se tornar um cientista de dados incluiu fazer um bootcamp, provavelmente estava fazendo sua busca de emprego enquanto desempregado. Se estava na faculdade, era esperado que tivesse tomado um tempo para fazer entrevistas e que passou esse tempo preparando seu currículo ou carta de apresentação (Capítulo 6). No entanto, se estiver trabalhando em tempo integral, seu gerente não quer ouvir que você precisa de uma folga porque está buscando um novo emprego. Então, como arranjar tempo para isso?

Coisas que podem ser feitas a qualquer momento – como atualizar seu currículo e redigir a carta de apresentação, buscar vagas, enviar currículos e fazer testes – devem ser realizadas no seu tempo livre. É melhor que as pessoas não vejam o que está fazendo, e você deve à sua empresa atual continuar fazendo bem seu trabalho. Mas as entrevistas acontecem quase



sempre durante suas horas normais de trabalho. Se as entrevistas forem realizadas por telefone, recomendamos que atenda à ligação em um espaço reservado, sala de reunião ou em outro lugar onde não possam ouvi-lo.

No entanto, as entrevistas de uma rodada posterior precisam ser feitas na empresa. Se uma entrevista tiver uma ou duas horas de duração e for perto de seu escritório, é possível dizer que vai ao médico. Se for mais longa, e sua empresa permitir que trabalhe de casa, você pode fazer isso e trabalhar apenas parte do dia (ou tentar trabalhar um pouco mais tarde), mas é preciso garantir que não esperem que você atenda a uma ligação ou responda rapidamente enquanto estiver em uma entrevista. Também é possível tentar agendar a entrevista para o fim do dia e trabalhar meio turno.

É muito mais fácil agendar entrevistas se estiver procurando trabalho na sua cidade. Se estiver buscando uma oportunidade de se mudar, a maioria das empresas fará uma entrevista presencial em um dia da semana. Nessa situação, é difícil evitar um dia inteiro de afastamento, sendo que a maioria das pessoas normalmente dirá que ficou doente naquele dia. Como é possível imaginar, é difícil fazer tudo isso se tiver de participar de muitas rodadas de entrevista.

Esta é uma das razões pelas quais recomendamos que se candidate de maneira estratégica a empregos. Se tiver dezenas de entrevistas por telefone e duas entrevistas locais em uma semana, é difícil arranjar tempo sem que as pessoas percebam, e o desempenho no seu trabalho será quase com certeza afetado de forma negativa. Você deve ser seletivo em duas etapas: ao se candidatar em primeiro lugar e ao avançar após a ligação inicial. Se estiver em uma startup e quiser trabalhar em uma empresa maior, não se candidate a outras startups, mesmo que as descrições das tarefas pareçam ótimas. Se, durante a entrevista inicial por telefone, descobrir que a função é mais na área de engenharia de dados e quiser trabalhar com análise, é bom suspender o processo mesmo que o entrevistador queira continuar.

Embora possa usar como prática entrevistas para funções que não aceitaria, não faça muitas dessa forma. Os cientistas de dados experientes são muito procurados, o que significa que você pode despertar o interesse de recrutadores e de gerentes ao anunciar que está buscando um novo emprego. É um ótimo sentimento: as pessoas gostam de você! Embora seja

bom desfrutar desse sentimento, não gaste tempo no processo de entrevista de uma empresa que você sabe que não se enquadra no seu perfil. Não é um bom uso do seu tempo, mesmo que seja lisonjeiro.

Na busca por emprego é fácil deixar seu trabalho atual decair. Quando você se sente bem em seguir adiante, muitas vezes fica pensando no que não gosta em seu emprego, o que pode minar a motivação. Tente continuar fazendo um bom trabalho; um dia você pode precisar de uma recomendação do seu gerente, além de a empresa ainda estar pagando seu salário.

É possível que, durante sua busca por emprego, perceba que a grama nem sempre é mais verde do outro lado. Em outras palavras, não encontrou algo que já não faça em seu emprego atual, ou as vagas pagam muito menos e oferecem benefícios piores, ou, ainda, não oferecem a flexibilidade de que desfruta atualmente. Tudo bem decidir interromper a busca por outro emprego! Não foi um desperdício de tempo se acabou percebendo que gosta do seu trabalho atual. Se decidir ficar, recomendamos que regresse ao nosso conselho na Seção 15.1.2: garanta ter tentado resolver quaisquer problemas que possa ter no seu trabalho atual.

## **Fazer pós-graduação**

Depois de trabalhar em ciência de dados, talvez você decida voltar à universidade para obter uma formação acadêmica mais formal, seja seguindo no trabalho em tempo integral e frequentando aulas à noites e nos fins de semana ou estudando em tempo integral. Se estiver pensando nisso, recomendamos voltar ao Capítulo 3, onde discutimos como encontrar um bom curso.

Queremos adverti-lo para que pense cuidadosamente se o investimento de tempo e dinheiro valerá a pena, embora já tenha provado que pode conseguir um trabalho em ciência de dados. Algumas razões pelas quais voltar à universidade pode fazer sentido: você decidiu que deseja uma função de pesquisa que requer doutorado, obteve feedbacks bem claros de empresas nas quais quer trabalhar de que é preciso ter mestrado (não só de ter visto na lista de requisitos das vagas), talvez percebeu que seu progresso está dificultado pela falta de determinadas competências (como conhecimento aprofundado de algoritmos) e as opções online e gratuitas não funcionam para você.

Se decidir voltar à universidade em tempo integral, dá para comunicar isso mais abertamente ao seu gerente do que quando quiser sair da empresa para outro emprego. Caso trabalhe em uma empresa maior, talvez possam até mesmo pagar parte do seu curso se estudar meio período e seguir trabalhando em tempo integral ou se concordar voltar em tempo integral após receber o título. Mesmo que não seja possível, seu gerente pode escrever uma boa recomendação. Um bom gerente sabe que a universidade oferece algo completamente diferente do trabalho e deve apoiá-lo em suas escolhas.

## **15.4 Como apresentar o aviso-prévio**

Se tiver decidido sair do seu emprego e aceitar uma proposta, terá de informar seu gerente. Nos EUA, em geral, deve-se dar pelo menos duas semanas de aviso prévio, a menos que a situação seja horrível. Embora improvável, é possível que assim que comunicar o aviso prévio, seu gerente o informe de que seu último dia será aquele. Prepare-se para essa possibilidade e garanta que tenha encaminhado documentos pessoais do computador do trabalho para seu computador pessoal.

Seu chefe deve ser a primeira pessoa a saber que você está saindo. Agende uma reunião com seu chefe (chame-a de “conversa sobre carreira” e não de “aviso prévio”) ou use o horário da reunião individual semanal. É melhor comunicar-se pessoalmente se trabalhar no mesmo escritório ou por telefone ou videochamada se não estiver; não avise por email. Comece a conversa expressando gratidão por como eles o ajudaram e pelas oportunidades que teve na empresa. Certifique-se de que fará tudo o que puder para ajudá-los na transição; é possível listar algumas ideias, como comentar seu código ou sugerir alguém para assumir uma parte de seu trabalho, mas pensar nisso tudo requer colaboração com seu gerente. É normal sentir-se ansioso para avisar ao seu gerente que está saindo, porém, lembre-se de que mudar de emprego é uma parte normal de uma carreira.

### **15.4.1 Considerando uma contraproposta**

É possível que seu gerente tente convencê-lo a permanecer na empresa com uma contraproposta, pois é oneroso e arriscado contratar um novo

funcionário. Podem lhe pedir que se reúna com um gerente de um nível mais alto, que tenha autoridade para oferecer um aumento, opções de ações extras, um bônus especial, uma promoção antecipada ou outros incentivos para que fique na empresa.

Não existe consenso sobre se você deve sempre aceitar uma contraproposta ou não feita por sua empresa atual. Por um lado, a empresa agora sabe que você pode sair a qualquer momento e pode relutar em oferecer-lhe mais responsabilidades. A situação também pode prejudicar a sua relação com o seu gerente. Por outro lado, a empresa pode estar disposta a abordar a razão principal pela qual quer sair. A solução poderia ser monetária ou poderia ser trocá-lo de equipe.

Esperamos tê-lo convencido da importância de manter uma comunicação aberta com seus gerentes. Se você tiver estabelecido um relacionamento de confiança com seu gerente e ainda sentir a necessidade de mudar para um novo emprego, é improvável que a mudança que você deseja possa ser efetivada em uma contraproposta. Embora não amemos a ideia de mudar para um novo emprego como uma última tentativa e sintamos que você deva tomar a decisão de sair bem antes desse ponto, salientamos a importância de comunicar seus desejos antes da sua insatisfação aumentar. Se tiver feito isso, tenha em mente que mudanças de última hora provavelmente não mudarão o ambiente geral de trabalho.

Seu gerente pode tentar destacar o valor que você traz para a equipe e como será difícil para a sua equipe se você sair. Isso pode fazer com que se sinta culpado, especialmente se seu chefe ou equipe em geral sejam do mesmo nível que você. Lembre-se, no entanto, de que não os está traindo por sair da empresa. No fim do dia, um trabalho é apenas um trabalho, e, apesar da retórica de algumas startups, uma empresa não é sua família. Embora deva sempre ser respeitoso e dar o melhor de si no trabalho, você não tem a responsabilidade de trabalhar nessa empresa por tempo indefinido. E você só está saindo de uma empresa, não morrendo! Se tiver se aproximado de seus colegas de trabalho, ainda poderá sair com eles e talvez até mesmo virem a trabalhar juntos um dia.

## **15.4.2 Como contar à equipe**

Fale para o seu gerente que gostaria de informar sua saída aos demais membros da equipe. Podem pedir-lhe que aguarde alguns dias enquanto pensam em um plano de transição para que possam compartilhar essas informações com a equipe quando comunicarem a sua demissão. Eles podem perguntar se você prefere contar a todos em uma reunião de equipe regular ou se quer se reunir com as pessoas individualmente. Recomendamos que considere o tamanho da sua equipe ao pensar nessa questão. Se trabalhou com as mesmas cinco pessoas de forma consistente ao longo dos anos, talvez queira contar a elas em particular. Por outro lado, se estiver em uma equipe de cientistas de dados de 20 pessoas e tiver trabalhado com stakeholders também, seria emocionalmente exaustivo reunir-se com todos eles individualmente dedicando meia hora a cada um.

Um erro a evitar é agendar reuniões com colegas antes de falar com seu gerente, mesmo que você programe essas reuniões para após a conversa com o gerente. Se seus colegas suspeitarem do motivo pelo qual, de repente, você quer reunir-se com eles e perguntarem se é porque está saindo, será realmente estranho: terá de mentir ou contar a eles antes do seu gerente saber.

A maioria das pessoas perguntará por que você está saindo. Pense no que dizer e tente manter-se positivo, concentrando-se na nova oportunidade e pelo que é grato na sua empresa atual. Mesmo que tenha se tornado amigo de um colega de trabalho, evite mesmo assim ser negativo. Lembre-se de que você pode querer voltar para essa empresa em algum momento, e é melhor não ter a reputação de alguém que falou mal da empresa ao sair. Algumas pessoas têm relações muito próximas com os gerentes, mas, mesmo que seja o caso, não fale muito sobre os aspectos negativos de seu trabalho atual, pois isso pode prejudicar sua amizade, além de ser desnecessário. Entre as outras boas razões para sair bem é que manter um bom relacionamento com seus colegas de trabalho e gerentes passados pode ser um ponto inestimável para a sua carreira, pois pode encontrá-los novamente ou precisar da recomendação deles mais tarde.

Ao se despedir, passe seu contato às pessoas (email, LinkedIn, Twitter e assim por diante). É interessante que os colegas de trabalho tenham como entrar em contato com você e também é uma boa forma de ser parte de uma

rede de contatos funcional para si e para os outros.

## **Checklist para observar antes de sair**

Antes de sair, certifique-se de que tem alguns itens administrativos:

- As informações de contato dos recursos humanos, caso necessite de algo mais tarde, como informações sobre suas opções de ações.
- Quaisquer imagens, senhas ou arquivos pessoais que tenha apenas no computador da sua empresa.
- Benefícios e informações de log-in do portal de remuneração de ações.
- Cópias de contratos de emprego, cartas de proposta e acordos de rescisão.
- Informações sobre como suas férias serão pagas se tiver algum período não usufruído.
- Se você não começar em um novo emprego imediatamente, suas opções para continuar no plano de saúde.
- Se tiver contribuído para uma poupança para cuidados de saúde ou de dependentes, qual é o último dia que pode gastá-la (geralmente seu último dia ou último dia do último mês de emprego). Esses fundos são daqueles de “se não usar, você perde”, portanto, se você não os utilizar, eles desaparecerão.

### **15.4.3 Como facilitar a transição**

A melhor maneira de sair bem é tornar a transição a mais fácil possível. Talvez você não consiga encontrar um substituto, mas pode ajudar a equipe a se ajustar enquanto estiverem buscando alguém para colocar no seu lugar (se preferirem isso). Faça um documento de transição para seu gerente, listando suas responsabilidades, quais consegue concluir, quais precisam ser transferidas (e sugestões sobre quem poderia assumi-las) e quais terão de esperar até que uma nova pessoa comece na função. Além de informar a pessoa que está assumindo esse projeto, você pode precisar fazer apresentações a parceiros externos ou a clientes, ou, ainda, informá-los de que não trabalhará mais no projeto.

Tente resolver tudo antes de sair. Se tiver algum trabalho que possa ser útil a outros, mas que esteja apenas no seu computador, adicione-o a um repositório git ou compartilhe o Google Doc com outra pessoa. Provavelmente, nessas últimas semanas, não receberá muito trabalho, porque as pessoas saberão que logo sairá. Assim terá tempo para fazer



coisas que antes não conseguia por estar com trabalho demais, como documentar todos os processos que você criou. Outras coisas que poderiam ser feitas incluem:

- *Adicionar tutoriais* – você era a pessoa a quem recorriam para um determinado tópico, por exemplo, como os dados financeiros são organizados ou as melhores práticas para testes A/B? Não há como substituir sua presença, mas, ao fazer apresentações, publicações internas ou documentação, dá para ajudar a preencher algumas das lacunas que deixará.
- *Organize seus arquivos* – mesmo que você adicione tudo ao GitHub, isso não ajudará ninguém se “tudo” for 100 arquivos com nomes como coisas\_aleatórias e análises\_diversas. Embora possa acabar precisando colocar alguns arquivos em uma pasta de arquivos extras, tente facilitar a navegação pelos arquivos e adicione explicações quando necessário.
- *Adicione comentários e explicações a análises* – em teoria, para quaisquer análises com impacto, você já escreveu os resultados, conectando ao código comentado. Se não tiver tido tempo para concluir alguns achados e acreditar que seria valioso para alguém continuar o trabalho, dá para colocar isso nos comentários. Embora não seja necessário comentar cada código, pode ser útil explicar algumas surpresas nos dados (e como você trabalhou com elas), o que você já tentou e explicar por que escolheu métodos analíticos.

A pior coisa a fazer é esquecer de que você é a única pessoa que sabe fazer *x* na empresa e não tratar desse assunto. Se não resolver isso, poderá receber ligações e emails impacientes sobre como fazer o trabalho enquanto tenta se adaptar no seu novo emprego. Esquecer que você possui a única senha para um determinado sistema é uma boa maneira de enfrentar problemas mesmo depois de sair. Alguns empregadores não sabem como dizer “adeus e boa sorte” e podem ligar para você a respeito de projetos nos quais você estava trabalhando. Para sua sanidade, é melhor se referir à sua documentação de saída até que eles entendam que você não trabalha mais lá. Mesmo que eles não façam isso, você não encontrará muito valor na rede de contatos que formou lá, caso deixe uma grande confusão para eles lidarem sozinhos.

Esperamos que este capítulo tenha deixado claro que, embora a incerteza sobre sair de seu emprego possa ser estressante, o processo é normal e há maneiras de torná-lo mais tranquilo. Como discutimos muitas vezes neste livro, poucas decisões que você toma são definitivas. Apenas porque começa a buscar outro emprego não significa que você tenha de sair, e, mesmo quando sair de uma empresa, pode acabar retornando mais tarde. O mais importante a ser feito durante todo esse processo é focar em fazer seu trabalho corresponder melhor aos seus objetivos de carreira.

## **15.5 Entrevista com Amanda Casari, gerente de engenharia da Google**

Amanda Casari é gerente de engenharia na Google, na equipe de Developers Relations do Google Cloud. Antes, foi gerente de produto principal e cientista de dados na SAP Concur. Ela também trabalhou cinco anos na Marinha dos Estados Unidos e tem mestrado em engenharia elétrica.

### **Quando se sabe que está na hora de buscar um novo emprego?**

Meu conselho para as pessoas é entender que tipo de trabalho querem fazer e se isso se enquadra na função e onde o produto, a equipe e a empresa se encontram. Por exemplo, dou-me muito bem em tempos de grande mudança. Gosto de trabalhar em projetos no início deles, durante a fase de idealização, mas também gosto de produtos na fase final. Por outro lado, não funcionaria muito bem para mim um trabalho de ciência de dados em que eu passaria a maior parte do tempo otimizando modelos para aumentos de porcentagem de um dígito ou fazendo ajustes de hiperparâmetros. Penso também na fase de coesão da equipe. Quer fazer parte de uma equipe que já tenha uma forte ligação e cultura ou de uma que está recém se formando? No geral, minha função, quando o produto está nesse ciclo de vida e quando a equipe está formando em comparação a estar em conformidade com uma cultura, influencia na decisão de se uma função ainda é boa ou não para mim ou se eu deveria procurar algo mais desafiador.

## **Já começou uma busca de emprego e acabou decidindo ficar?**

Quase sempre. Quando estou buscando outras funções, consigo identificar coisas que poderia estar realizando na minha empresa atual. É interessante encontrar essas oportunidades, em vez de esperar que elas caiam do céu. Isso se encaixa na minha filosofia pessoal de que as responsabilidades do seu atual trabalho devem ser sempre tópico de conversas com o seu gerente. Para os engenheiros que gerencio, tento ter conversas abertas e honestas quando me contam quais oportunidades estão buscando. Posso então descobrir se existe algo assim na equipe atual ou, se não, se podemos encontrar um projeto 20% fora da nossa equipe. Dessa forma, eles podem ter uma ideia se aquilo é ou não o que querem fazer.

## **Você vê pessoas permanecendo no mesmo emprego por tempo demais?**

Ah, sim. Já vi algumas pessoas com algum tipo de complexo de herói, onde sentem como se ninguém pudesse fazer o trabalho delas. A resposta real é que ninguém fará seu trabalho exatamente como você faz, mas não significa que ninguém mais consegue fazê-lo. Às vezes, é muito prejudicial alguém ficar muito tempo em uma equipe porque recordam todos os problemas e pequenas decisões pelos quais a equipe já tomou. Podem dizer: “Tentamos essa ideia dois anos atrás e não funcionou, não devemos tentar de novo”. Talvez acabe fazendo a equipe não se concentrar no que é possível no momento, mas só nas decisões tomadas no passado.

Também vi pessoas que ficaram bastante cansadas com a gestão e passam muito tempo se queixando. Esses funcionários estão sempre prontos para fofocas, o que é realmente negativo para uma empresa. Ninguém quer pessoas que transformam o descontentamento delas em algo que intoxica o restante da equipe.

Por fim, já vi pessoas que não se sentem desafiadas nas funções atuais e que fazem apenas o que lhes é solicitado e nada mais. Pode ser aceitável quando se é mais novo, mas para pessoas experientes e aquelas na liderança, espero mais do que isso. Gosto de ver pessoas experientes promovendo mais impacto e mudanças. Se você encontrar um problema, deve pensar em abordá-lo de uma forma que seja escalável, possível de

reproduzir e resolva um problema organizacional, não apenas algo que é uma solução única.

### **Dá para trocar de emprego com muita rapidez?**

Quando estou buscando candidatos, posso perguntar se ele teve um emprego que durou menos de um ano. Como gerente de contratação, o que você tenta entender ao observar a duração de uma pessoa em uma empresa é se elas vão se demitir em poucos meses, pois contratar e fazer a integração de alguém é um processo longo e caro. Embora em outros setores seja possível ver de dois a três anos como o mínimo, na tecnologia, considero esse período como um tempo longo. Em dois a três anos é possível construir alguns projetos com vários ciclos em tecnologia, por isso, sair por esse período faz sentido para mim; na verdade, mais de um ano já faz sentido. Se tiver que sair antes de um ano por conta da sua saúde mental ou emocional, faz muito sentido também; por nenhum trabalho vale a pena colocar isso em risco.

### **Qual é seu último conselho para os cientistas de dados iniciantes e aspirantes?**

Encontre sua comunidade e pessoas que podem ajudá-lo. Já me beneficiei tanto por ter um amigo querido que me patrocinou ao me recomendar para falar sobre oportunidades e me passar ofertas de emprego. Ter uma pessoa com experiência com quem falar sobre esses detalhes é inestimável e realmente ajudou-me a entender meu valor e a me sentir confiante ao começar em um novo cargo. Para encontrar uma comunidade, há tantos lugares aos quais pode pertencer, mas pode haver algumas nas quais não experimente esse sentimento de pertencimento. Você não precisa ficar em um lugar onde se sinta desconfortável, seja por causa da linguagem, das pessoas que estão ali reunidas ou do foco do grupo. E se não conseguir encontrar um espaço em que se sinta confortável, procure pessoas que queiram estar nesse tipo de grupo e pergunte se podem ajudá-lo a formar um.

## Resumo

- Ao decidir se deve buscar um novo emprego, há quatro perguntas a fazer: se ainda está aprendendo, se falou com seu gerente para ver se suas responsabilidades poderiam mudar, se os objetivos de carreira do seu gerente estão alinhados com os seus e se você pensou no que está buscando (e naquilo que não) em sua próxima função.
- Embora muitos princípios para ter uma primeira busca exitosa de trabalho na ciência de dados (que você pode encontrar na Parte II deste livro) ainda se apliquem, para seu segundo emprego, reflita também sobre o que você gostou (e não gostou) do seu primeiro emprego. Prepare-se para compartilhar sua experiência de uma maneira positiva, que respeite a confidencialidade da empresa e planeje como encontrar tempo para entrevistas com seu emprego em tempo integral.
- Após ter apresentado o aviso-prévio ao seu gerente, concentre-se em como tornar a transição o mais fácil possível para seus colegas de equipe, deixando tudo resolvido, documentando tudo o que atualmente está só na sua cabeça e compartilhando algum código útil.

# CAPÍTULO 16

## Como subir na carreira

Este capítulo abrange:

- Caminhos diferentes para além de cientista de dados sênior
- As oportunidades e riscos de possíveis trajetórias na carreira

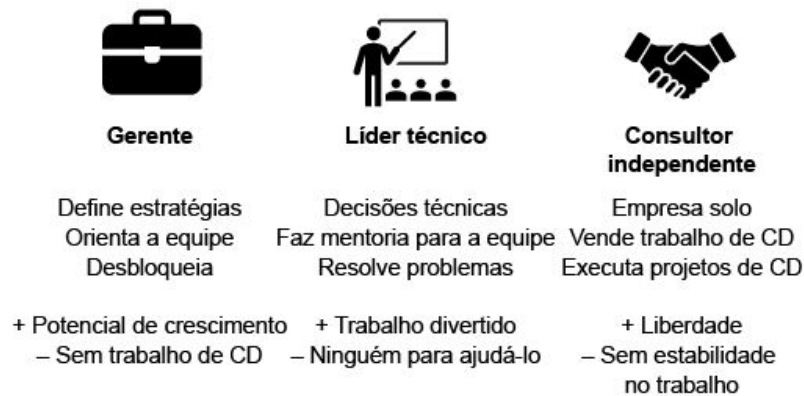
Na última parte deste livro, falamos sobre como desenvolver mais sua carreira: aprender a lidar com o fracasso, participar da comunidade e mudar de emprego. À medida que sua carreira se fortalece, você acabará decidindo para onde ela irá. Não é óbvio quais opções os cientistas de dados terão à medida que galgam postos mais elevados na carreira; tornar-se gerente é uma possibilidade, mas não é a única.

Neste capítulo, abordaremos de três trajetórias de carreira comuns para um cientista de dados – ir para a gerência, tornar-se um líder técnico e mudar para uma consultoria independente – bem como os benefícios e desvantagens de cada uma.

A opção da gerência é o que a maioria das pessoas imagina quando pensa em crescimento na carreira. *Gerentes* são as pessoas que lideram equipes, incluindo contratação e promoção, definição de estratégia e orientação de carreira. *Cientistas de dados líderes* são os mestres na área, e as empresas dependem deles para resolver problemas técnicos difíceis. *Consultores independentes* são cientistas de dados que têm competências suficientes e uma rede de contatos grande o bastante para serem capazes de fazer trabalhos como freelancer. Consulte a Figura 16.1 para ver um resumo dos caminhos neste capítulo.

Ao seguir moldando sua carreira, é bom se concentrar em um desses caminhos. Tendo um objetivo claro, é mais provável conseguir o que quer. Dito isso, quanto mais perto chegar de uma oportunidade desejada, mais poderá perceber que, no fim, não é algo que você quer. Felizmente, fazer

mudanças para se tornar um gerente e depois decidir que não gosta disso ou sair do trabalho formal para consultoria e depois voltar novamente para o setor, não é incomum. E, embora reverter as decisões que tomou possa ser difícil, aprender com seus erros é uma maneira rápida de você crescer como pessoa!



*Figura 16.1 – Os caminhos neste capítulo.*

**Como saber em qual nível você se encontra**

Conforme você cresce como cientista de dados, mais competências valiosas ganhará (como discutido nos capítulos anteriores). Mas, à medida que seu conjunto de competências e sua maturidade profissional crescem, em algum momento você não estará mais trabalhando como cientista de dados júnior. É difícil saber exatamente quando essa transição acontece, o que torna difícil decidir o momento certo para pedir uma promoção ou um novo cargo. Tenha em mente que cada empresa tem seus próprios níveis e expectativas; por isso, o mesmo título pode significar coisas muito diferentes em duas empresas distintas. Em uma empresa pode até haver uma matriz de competências descrevendo exatamente como os níveis se diferem, mas as pessoas podem interpretar a matriz de forma diferente, e sempre há ambiguidade. Para ajudá-lo, eis nosso guia de alto nível sobre as diferentes expectativas para níveis de cientistas de dados:

- *Cientista de dados júnior* – uma pessoa que pode completar uma tarefa de ciência de dados a partir de uma orientação clara sobre como essa tarefa deveria ser. Se lhe for solicitado usar um algoritmo de cluster para segmentar novos clientes por seus atributos de compra, um cientista de dados júnior deve ser capaz de fazer isso com a orientação de um gerente. Se ocorrerem problemas técnicos, como bugs em sistemas de código ou dados que não se conectam, eles podem precisar consultar outros membros da equipe para resolvê-los.
- *Cientista de dados sênior* – uma pessoa que não só pode concluir uma tarefa de ciência de dados, como também descobrir quais outras tarefas são necessárias. Não só poderiam construir o exemplo de segmentação com cluster (descrito anteriormente), mas igualmente perceberiam coisas como o fato de que o mesmo algoritmo poderia ser usado também para clientes existentes (e, então, fazer isso). Eles são qualificados para resolver problemas técnicos e acionados quando outros têm problemas.
- *Acima do cientista de dados sênior* – em níveis mais altos do que o cientista de dados sênior, a função torna-se mais uma questão de ajudar os outros. Assim, uma pessoa geralmente vai além do cargo de cientista de dados sênior quando está fazendo a mentoria de outros, criando estratégias e analisando o panorama geral.



## 16.1 A via da gerência

Ingressar na gerência pode parecer ser a opção padrão para os cientistas de dados à medida que vão mais longe em suas carreiras. Faz sentido devido à exposição: todos têm uma pessoa a quem se reportam. Apesar desse fato, as tarefas diárias de um gerente ainda podem ser desconhecidas.

Um gerente é alguém responsável por uma equipe que executa com sucesso os próprios objetivos. As cinco tarefas básicas, apresentadas a seguir são normalmente, mas nem sempre, o trabalho de um gerente de ciência de dados:

- *Definir o trabalho da equipe* – esse trabalho pode ser feito em nível estratégico, como, por exemplo, decidir quais projetos grandes deve assumir. Também pode ser feito em um nível mais tático, como decidir quais recursos devem ser incluídos em um produto.
- *Determinar quem deve integrar a equipe* – um gerente via de regra escolhe quem contratar e quem demitir, com a aprovação dos recursos humanos e de outras pessoas. Ele coordena o processo da entrevista e opina durante a seleção.
- *Fazer a mentoria da equipe* – todas as pessoas de uma equipe têm desafios únicos para trabalhar e um gerente que as ajuda a enfrentar esses desafios. Um gerente verifica regularmente cada pessoa e oferece conselhos e recomendações para ajudá-las a resolver os problemas.
- *Resolver problemas da equipe* – se a equipe estiver com problemas que impeçam o trabalho a ser feito (por exemplo, se outra equipe não estiver disposta a fornecer o acesso necessário aos dados), é tarefa do gerente encontrar uma solução para que a equipe possa continuar.
- *Gerenciar projetos* – um gerente precisa acompanhar o trabalho que está sendo realizado pela equipe e garantir que as coisas estejam dentro do cronograma. Embora muitas equipes tenham um gerente de projeto especificamente designado para cada tarefa, ele ainda precisa manter a supervisão do trabalho.

Em conjunto, essas tarefas abrangem uma vasta gama de trabalho e envolvem uma elevada responsabilidade. Um gerente tem de se comunicar

continuamente com pessoas dentro e fora da sua equipe. Também precisa estar ciente de como os integrantes da equipe estão indo, como os projetos estão andando e o que há no horizonte. Essas tarefas somam uma enorme quantidade de trabalho e, se alguma delas não forem bem realizada, toda a equipe será atingida.

É importante notar que as tarefas básicas de um gerente não são técnicas; um gerente geralmente não está criando modelos de machine learning ou fornecendo análises que ajudam a empresa a tomar decisões. Um gerente não tem tempo para fazer esse tipo de trabalho e, mesmo que o fizesse, seria melhor um cientista de dados fazê-lo. Tornar-se gerente envolve desistir de muito do que motiva as pessoas a se tornarem cientistas de dados em primeiro lugar: usar dados para resolver problemas interessantes. Em vez disso, o trabalho é apoiar outras pessoas que fazem esse trabalho.

### **16.1.1 Vantagens de ser gerente**

Ser gerente proporciona muitas vantagens. Primeiro, se você for uma pessoa que odeia quando outras fazem um trabalho medíocre, ser gerente significa que você começa a influenciar as situações para que isso não aconteça. Se achar que um modelo de machine learning em particular seria útil para a empresa, por exemplo, você pode designar a equipe para fazer isso, e se acha que não é uma boa ideia, pode garantir que a equipe evite isso. É muito gratificante orientar a equipe e, em seguida, ver esse grupo bem-sucedido. Apesar de você não ter controle total – por vezes, as pessoas em uma posição superior exigem que algo seja feito ou as pessoas da sua equipe recomendam incisivamente que alguma coisa deve ser feita – sua opinião tem peso.

Alcançar o nível de gerência muitas vezes significa um aumento inicial da remuneração comparado ao de colaborador individual. A gerência também abre portas para outras funções: gerente sênior, diretor ou vice-presidente. Cada uma dessas funções tem um salário mais elevado e uma liderança maior na empresa. Você pode até chegar a um nível no qual supervisiona mais áreas do que apenas a ciência de dados, como pesquisa de cliente ou desenvolvimento de software. Por fim, é possível fazer a transição total para fora da ciência de dados e liderar outras áreas.

Se você gostar de ensinar e ajudar os outros, a gerência é uma boa função. Grande parte do seu trabalho será ajudar a sua equipe a se tornar bem-sucedida trabalhando com seus integrantes, ensinando-lhes o que aprendeu em sua carreira e ajudando-os quando encontram dificuldades. Um bom gerente é como um terapeuta corporativo: senta-se com uma pessoa por 60 minutos e a ajuda a lidar com os problemas.

Além disso, você pode ter uma esfera imensa de influência como gerente. Ser a pessoa que toma as decisões finais com relação a criar ou não um novo produto ou se expandir para um novo país é realmente interessante! À medida que sobe degraus na empresa, sua influência poderá aumentar cada vez mais. Se seguir esse caminho, terá a possibilidade de um dia administrar uma empresa.

### **16.1.2 Desvantagens de ser gerente**

A maior desvantagem de ser gerente é o fato de não ter tempo para fazer ciência de dados. Seu trabalho será repleto de coisas tais como falar sobre ciência de dados, fazer a mentoria de cientistas de dados e pensar sobre estratégias de ciência de dados, mas o gerente mesmo não fará ciência de dados. Seu dia pode ser preenchido com reuniões de 30 minutos que vão desde a decisão da estratégia da equipe a agradar stakeholders para conseguir recursos até reuniões individuais para ajudar funcionários com um desempenho inferior a aprenderem a como melhorar.

Não fazer mais ciência de dados como seu trabalho é uma desvantagem por duas razões:

- Provavelmente tornou-se um cientista de dados por gostar de trabalhar com dados, então, está desistindo do trabalho para o qual treinou durante tanto tempo.
- Enquanto não fizer ciência de dados, você ficará sem prática e afastado das mudanças mais recentes. Se decidir que não gosta de gerenciamento e quiser voltar a ser um colaborador individual, talvez perceba que suas competências de ciência de dados diminuiriam.

Outra desvantagem é que ainda fica limitado pela gerência superior à sua. É possível ter grandes ideias para a estratégia da sua equipe que são refutadas

pelo seu chefe. Ainda é preciso liderar sua equipe no caminho definido pelo seu chefe, mesmo que não concorde com ele. Ter que manter uma atitude positiva sobre o trabalho com o qual não concorda para o benefício de sua equipe pode ser extremamente frustrante. Se deixar seus subordinados sentirem suas frustrações, é capaz também de baixar a satisfação deles no trabalho.

Ser gerente é ter muito mais coisas para se preocupar. Tem de preocupar-se com seu próprio desempenho, tal como quando colaborava individualmente, mas também com o desempenho do resto da sua equipe. Tem de preocupar-se com a satisfação da sua carreira e com o que está acontecendo politicamente nos níveis superiores ao seu. Tem de se preocupar se a equipe terá recursos e se um projeto que está se desenvolvendo muito devagar será cancelado. Ter tantas preocupações, e a maioria delas não estarem diretamente sob seu controle, pode ser muito estressante para pessoas com certos tipos de personalidades. Se tiver problema em levar trabalho para casa, a gerência talvez não seja para você.

Finalmente, gerenciar pessoas requer um conjunto totalmente diferente de habilidades do que ser um cientista de dados. Quando chegar a gerente, terá de aprender essas novas competências e voltar a ser um principiante no seu trabalho. Essa virada de deixar de ser ótimo em seu trabalho para ser um novato pode ser estressante e triste. Embora finalmente aprenda a função, a trajetória para se tornar um bom gerente é longa.

### **16.1.3 Como se tornar um gerente**

Se você for um colaborador individual e almejar a gerência, precisa encontrar oportunidades para desenvolver e praticar competências de liderança, as quais incluem trabalhar bem com outras pessoas, tanto mais iniciantes quanto mais experientes do que você, ver o panorama geral e gerenciar um cronograma para um projeto.

Infelizmente, não há um único caminho a tomar para aprender essas competências; o melhor a fazer é encontrar situações em seu trabalho atual onde possa crescer. Um exemplo é uma pequena iniciativa na sua equipe que alguém precisa assumir, como configurar um novo software ou coordenar a implantação de um novo modelo. O componente importante

dessas situações é que você é o líder e toma as decisões. Pode parecer extremamente artificial no início, mas esse sentimento é completamente normal e diminuirá com o tempo. Os livros sobre gerência e administração podem ser úteis, mas somente se tiver oportunidade de usar o que aprendeu.

Quando sentir que aprendeu as competências para ser gerente, a etapa seguinte é encontrar uma função.

## **Conseguir uma promoção na empresa**

Muitas vezes, a forma mais simples de se tornar gerente em sua empresa é receber uma promoção, o que pode ser mais fácil, pois as pessoas com poder de decisão para colocá-lo na função de gerente são as que viram suas competências melhorarem no trabalho atual. A dificuldade com essa rota é que a empresa tem de precisar de um novo gerente: seu gerente atual tem de sair ou ser promovido, ou uma função de gerente precisa ser aberta em uma equipe relacionada. Dependendo da empresa, essas situações podem raramente ocorrer ou talvez nem ocorram.

## **Como formar uma nova equipe você mesmo**

Outra rota é tornar-se um gerente que forma sua própria uma equipe. Talvez comece um projeto em sua empresa atual que acaba precisando de mais pessoas ou se torne o primeiro cientista de dados em uma nova empresa e, depois, forma ali uma equipe. Esse caminho pode ser extremamente recompensador, porque você mesmo formar uma equipe permite-lhe ter mais controle sobre quem a integra e como funciona. O trabalho exige que você esteja no lugar certo e no momento certo, além de ser um líder forte o suficiente para ter condições de aumentar com rapidez toda a infraestrutura da equipe. Por vezes, as pessoas que tomam esse caminho são chamadas de jogador-treinador porque trabalham tanto como o primeiro colaborador individual da equipe quanto se tornam treinador para os outros. Infelizmente, as oportunidades para criar sua própria equipe são ainda mais raras do que o caminho mencionado anteriormente.

## **Conseguir a função de gerente em uma nova empresa**

A última abordagem é ser contratado para uma vaga aberta de gerente em

uma empresa diferente da sua. Esse caminho depende de sua capacidade de mostrar próprias competências de gerenciamento sem ter sido tecnicamente um gerente antes. Na sua carta de apresentação e ao fazer a entrevista, terá de falar sobre todos os projetos que lidera e sobre as pessoas a quem fez mentoria como colaborador individual. Pela quantidade de atenção que seu currículo receber ao compartilhá-lo, você consegue dizer se tem chance de seguir esse caminho; as empresas estão desesperadas por bons gerentes de ciência de dados; então, se seu currículo parece bom, deverá receber retorno.

### **Rob Stamm, diretor que supervisiona a equipe de IA na T-Mobile: como aprender a gerenciar**

Aprendi muito sobre a atividade de gerente durante minha primeira experiência em uma gerência em tempo integral. Antes disso, fui gerente de produto, que significava orientar a direção e o desenvolvimento de um produto (mas não as pessoas). Então, como gerente sênior recém-contratado, tive de supervisionar vários gerentes de produtos diferentes e ajudá-los a fazerem seu trabalho. Acabei fracassando nessa função. Queria ser gerente sênior e liderar uma equipe, mas não conseguia deixar a função de gerenciamento de produto também. Não deixava a equipe fazer o trabalho dela; continuei a fazer por eles.

Quatro meses no trabalho, um dos gerentes de produto entrou na minha sala e disse-me que eles iriam embora e que a razão era minha maneira de agir. Foi uma grande constatação para mim ver que não dá para ser gerente e continuar tentando ser um colaborador individual. Foi a primeira vez que alguém apontou que um gerente tem de permitir que sua equipe tome decisões, então levei isso comigo para sempre.

Por fim, aprendi e cresci com essa experiência e comecei a liderar equipes maiores e projetos também maiores. Ser um líder é incrivelmente gratificante: tenho o trabalho de ajudar minha equipe a ser a melhor e me sinto satisfeito quando vejo os resultados. Um projeto em particular na T-Mobile especialmente gratificante foi conseguir ajudar a equipe de IA a começar apenas a partir de uma ideia e alguns dólares de financiamento até chegar a uma equipe de grande escala. Embora não estivesse concebendo o produto ou escrevendo o código, auxiliei a equipe quando estava se sentindo travada ou necessitava de recursos, como dinheiro ou pessoas. É gratificante à sua maneira!

## 16.2 O caminho para ser cientista de dados líder

Um cientista de dados líder (ou da equipe, chefe, cientista de dados V, chefe técnico ou outro título, dependendo da empresa) é alguém cujo trabalho é ser um especialista em ciência de dados e ajudar outros com tarefas técnicas. Tornar-se um gerente envolve fazer cada vez menos ciência de dados, enquanto ser um cientista de dados líder permite fazer. Mas, em vez de fazer ciência de dados sozinho (embora faça muito isso), também terá a tarefa de ajudar outros cientistas de dados nos respectivos trabalhos.

Os cientistas de dados líderes geralmente começam como cientista de dados júnior, são promovidos a cientista de dados sênior e continuam crescendo para além dessas funções. À medida que crescem nas suas carreiras de ciência de dados, tornam-se mais experientes e maduros na capacidade de compreender problemas e resolvê-los. Logo, sabem tanto que quando outras pessoas estão apresentando dificuldades, eles conseguem ajudá-las de imediato. Pessoas de toda a empresa vão até eles pedir-lhes sugestões sobre como lidar com problemas e o que provavelmente funcionaria (ou não).

Um trabalho de cientista de dados líder envolve várias responsabilidades:

- *Influenciar a estratégia de ciência de dados* – um cientista de dados líder tem de definir um plano para resolver problemas de ciência de dados. A modelagem de fraude de pagamento é viável? Deve ser utilizada uma rede neural? O gerente é responsável pela ideia e pelo plano de negócios; o líder é responsável pelo funcionamento.



- *Fazer a mentoria de cientistas de dados júnior* – como o cientista de dados líder tem muita experiência, é sua obrigação compartilhar o que sabe com funcionários menos experientes. O crescimento deles é tão importante quanto o trabalho dos próprios cientistas de dados líderes.
- *Encontrar soluções para problemas difíceis* – quando a equipe de ciência de dados está sendo posta à prova por um problema técnico complexo, o cientista de dados líder é o responsável por elaborar uma solução – ou declara que é algo impossível de resolver.

Comparado com um gerente, um líder ainda tem muito trabalho de ciência de dados. Portanto, essa função é ótima para pessoas que amam ser cientista de dados. Se tudo o que consegue pensar é ciência de dados, gostar de ir a conferências, fazer parte da comunidade e aprender mais sobre técnicas e métodos, continuará fazendo tudo isso como cientista de dados líder. Devido à importância que um cientista de dados líder pode ter em uma equipe, a função requer uma pessoa com maturidade, responsabilidade e experiência suficientes em ciência de dados que aprenda rapidamente sobre novas áreas.

Embora o trabalho de um cientista de dados líder esteja relacionado à ciência de dados, é raro que um líder realize trabalhos de colaborador individual, como fazer uma única análise ou criar um modelo de machine learning. Essas tarefas exigem muito tempo e foco unificado, e um cientista de dados líder tem de dividir seu trabalho em muitos projetos e áreas. Os projetos de ciência de dados tendem a ficar com cientistas de dados júnior e sênior, embora o líder os coordene e supervisione.

## **Pedir uma promoção**

Em algum momento, provavelmente se sentirá pronto para conquistar outro nível, mas não terá sido promovido. Essa situação pode ser frustrante (você pode entender porque está pronto; por que ninguém vê!?), mas não é impossível. O melhor para ajudar a avançar nesse processo é ir atrás disso. Informe seu gerente de que está interessado em passar a outro nível e que pretende trabalhar com seu gerente para criar um plano para alcançar seu objetivo. Se seu gerente for bom, deve ficar satisfeito com isso: ao deixar claro que está pronto para fazer mudanças para alcançar o próximo cargo, a conversa foi iniciada sobre como chegar lá. Tente definir um objetivo com uma data específica, como o próximo ciclo de análise de desempenho para ser promovido. Os objetivos devem ser tão específicos quanto possível, como “realizar três apresentações técnicas na empresa” ou “criar e implantar uma API de machine learning inteira por conta própria”. Ao ter objetivos claros com um cronograma, você poderá ter informações regulares sobre seu progresso.

Se seu gerente oferecer feedback sobre o motivo pelo qual não está pronto, escute-o. Embora seja difícil ouvir um feedback negativo, seu gerente tem uma perspectiva que você não tem sobre o que precisa para a promoção. Se seu gerente falar que você está pronto e que pretende ajudá-lo no processo de promoção, tente preparar o máximo de documentação possível sobre o que tem feito e por que razão é capaz de assumir o novo trabalho. Essa documentação pode ajudar seu gerente.

Se parecer que não importa o que você faz, ainda pode não conseguir a promoção que espera, talvez seja um sinal de que é hora de ir para outra empresa. Há muitas ocasiões nas quais as pessoas estão tão habituadas a ver alguém em uma determinada função que não estão dispostas a arriscar colocar essa pessoa em uma nova. Ao mudar de empresa, você começará a trabalhar com um grupo de pessoas sem expectativas sobre o que você pode fazer e podem oferecer-lhe mais oportunidades.

### **16.2.1 Benefícios de ser um cientista de dados líder**

Um cientista de dados líder recebe com frequência as atividades mais interessantes. Se a equipe tiver uma ideia sobre abordagem totalmente nova de ciência de dados que ninguém tenha tentado antes, você estará envolvido nessa tentativa pela primeira vez. Se um recurso técnico complexo precisar ser integrado, você estará lá. Se um projeto deveria estar sendo concluído, mas a equipe não consegue fazer o modelo funcionar, você também estará lá. Estar no meio dessa ação toda pode ser muito compensador. É interessante pelo aspecto técnico, e você começa a se sentir validado à medida que é útil inúmeras vezes. A equipe estará ciente de que projetos simples e amadores de ciência de dados não valem seu tempo tão limitado, o que significa que terá de atuar menos nesse tipo de tarefa do que a maioria dos cientistas de dados.

Seu gerente também compreenderá quão importante é manter-se a par da tecnologia, o que significa que provavelmente terá financiamento para ir às conferências e para entreter-se com a nova tecnologia. Dependendo do seu gerente, você pode se apresentar em conferências para promover sua empresa, mas, como atua em problemas interessantes, provavelmente terá feito algum trabalho que vale a pena ser compartilhado. Se pedir recursos ao seu gerente, como dinheiro para experimentar um novo serviço em nuvem, normalmente conseguirá. Seu gerente confiará em você para usar seu tempo e orçamento de forma eficaz e não desperdiçá-lo, que é algo que nem todos os cientistas de dados conseguem.

Do mesmo modo, poderá falar sobre ciência de dados o tempo todo. Como fará a mentoria de cientistas de dados iniciantes, poderá contar a eles sobre diferentes abordagens, trabalhar com eles para aperfeiçoar as ideias

deles e indicar áreas onde as abordagens de ciência de dados podem apresentar problemas. Para uma pessoa que gosta de ciência de dados, essa função pode ser bastante interessante.

Elaborar planos de ciência de dados pode ser muito gratificante. Ser quem decide que tipos de modelos usar, como estruturar dados e como fazer o escopo de um projeto, é mais provável que veja os projetos feitos da maneira que deseja. E como você é um especialista em ciência de dados, os projetos devem ter mais probabilidade de serem bem-sucedidos! Os cientistas de dados iniciantes nem sempre conseguem conceber uma abordagem de ciência de dados, executá-la e ver os próprios resultados sem que outra pessoa tome decisões fundamentais sobre ela.

### **16.2.2 Desvantagens de ser um cientista de dados líder**

O maior problema de ser um cientista de dados líder é que você não terá a quem recorrer quando estiver se sentindo travado. Um cientista de dados júnior geralmente tem um mentor ou um cientista de dados sênior a quem fazer perguntas, ou pode até mesmo fazer buscas no Google em busca de respostas aos seus problemas. Como cientista de dados líder, provavelmente você não terá ninguém mais experiente do que você em sua equipe. Os problemas que enfrentará muitas vezes serão tão incomuns ou únicos que ninguém os enfrentou antes, então, nenhuma pesquisa no Google fornecerá uma resposta. Assim, precisará ser capaz de trabalhar em ambientes nos quais não tenha ajuda, o que pode ser devastador para muitos profissionais.

Embora possa estar envolvido nos problemas mais relevantes da ciência de dados, também pode enfrentar os mais desinteressantes. Se, por exemplo, um conjunto de dados for armazenado como terabytes de arquivos .csv mal formatados em um servidor que ninguém toca há anos e com um esquema (schema) que ninguém conhece, você será chamado para descobrir como usar os dados. Esse problema não é realmente interessante; é apenas um campo minado pelo qual ninguém mais consegue andar. Você enfrentará muitos problemas como esse e, provavelmente, não poderá delegar as soluções a outras pessoas.

Como seu conhecimento é tão vasto, você estará em alta demanda – e sempre muito ocupado. Frequentemente terá mais trabalho do que tempo

para fazê-lo e terá de deixar de lado projetos interessantes porque você não tem tempo para eles. É fácil cair no hábito de trabalhar mais horas do que deve e ainda sentir que não fez o suficiente. Como tantas pessoas dependem de você, é bem mais difícil trabalhar apenas 40 horas semanais ou ter férias longas sem seu computador. Se quiser um trabalho agradável e relaxante, ser cientista de dados líder provavelmente não será assim.

### **16.2.3 Como se tornar um cientista de dados líder**

Se você trabalhar como cientista de dados e continuar recebendo promoções, o padrão é se tornar um cientista de dados líder. A função é a progressão natural do cientista de dados sênior; é uma versão ampliada dessa função. Infelizmente, muitos cientistas de dados sênior lutam para serem promovidos. A fim de se qualificar para ser um cientista de dados líder, você precisa ser um cientista de dados bom o suficiente para trabalhar de forma independente e liderar os outros. Também precisa chamar a atenção para suas capacidades e contribuições e encontrar outras pessoas na empresa que o sigam, para que seja bem conhecido como parte fundamental da equipe. Com essas qualificações, é possível pedir uma promoção para essa função mais elevada.

Para trabalhar de forma independente como cientista de dados, você precisa ser capaz de lidar com projetos completos sem orientação externa. Se seu gerente atribuir uma tarefa como “criar uma análise sobre onde devemos abrir nossa próxima loja de varejo”, eles precisam ser capazes de confiar que você consegue fazer o trabalho sem a ajuda de outros, sem ficar travado e não dizer a ninguém. À medida que cresce como cientista de dados, a capacidade de trabalhar com independência deve vir naturalmente à medida que se torna mais experiente no trabalho. Para acelerar esse crescimento, tente prestar atenção a quando fica travado e o que faz nessas situações. Se pedir ajuda, o que essa ajuda oferece que não conseguiu sozinho? Quanto mais conseguir maximizar os momentos em que pode resolver o problema por si mesmo, melhor.

À medida que ganha experiência, também preste atenção aos cientistas de dados ao seu redor. Quais são os problemas deles? E você é capaz de ajudá-los? Se você for um cientista de dados mais sênior, é provável que os

funcionários júnior estejam lidando com problemas que você enfrentou e resolveu no passado. Quanto mais oportunidades puder encontrar para a mentoria técnica, mais forte ficará em ajudar os outros – o que é ótimo para a função de líder.

Por fim, como novas ideias são lançadas, procure situações nas quais pode criar uma abordagem. Se a empresa quiser encontrar locais para lojas de varejo, por exemplo, você poderia ter a ideia de usar técnicas de otimização de localização. Suas ideias podem ou não funcionar, mas, se funcionarem, você se dará muito bem; se não funcionarem, aprenderá mais sobre ter ideias. É fácil contar com outros cientistas de dados para formular uma abordagem, mas é difícil, em contrapartida, ser um cientista de dados líder sem essa competência.

## 16.3 Mudar para consultoria independente

É um sonho comum ser seu próprio chefe e ter sua própria empresa. No caso da ciência de dados, em geral significa ser um consultor independente: ter um negócio que as empresas contratam para trabalhar em projetos de ciência de dados especializados. Em teoria, as empresas querem empregar um consultor externo somente se necessitarem de um conjunto especial de competências para um problema importante. Se você estiver administrando sua própria empresa, poderá manter todas as receitas para si mesmo, de modo que nenhum dinheiro será perdido com os executivos de altos salários que estão acima de você. As pessoas irão contratá-lo porque acreditam que você é um ótimo cientista de dados e, então, será valorizado pela sua especialidade.

Como consultor independente, você tem de ser uma empresa inteira, o que significa que tem de fazer muitas coisas diferentes, como:

- *Fazer o marketing da sua empresa* – você não conseguirá novos clientes para sua empresa a menos que as pessoas o conheçam. Pode significar ir a conferências, ter reuniões com antigos colegas ou criar materiais promocionais, como posts de blog.
- *Fazer vendas* – quando encontrar uma empresa interessada em contratar sua empresa para um projeto de ciência de dados, precisará se reunir

com pessoas dessa empresa e fazer uma proposta para o trabalho. Se não conseguir interessá-los por essa proposta, não conseguirá o trabalho.

- *Executar o projeto* – esse é o trabalho de ciência de dados para o qual você foi contratado. Também a gerência de projetos para manter as coisas em um bom caminho e lidar com quaisquer situações que possam surgir (como dados inválidos).
- *Apresentar resultados* – após criar o modelo ou fazer a análise, você deve apresentar o que fez ao cliente e colocá-lo a par de tudo. Se gostarem, talvez consiga outro projeto, mas, se não gostarem, poderá perdê-los como cliente.
- *Gerenciar a empresa* – empresas precisam fazer coisas como pagar impostos, criar documentos legais, manter registro das contas e do fluxo de caixa, além de executar muitas outras pequenas tarefas que vão sendo adicionadas com o tempo.

Você precisa de competências para concluir todas essas tarefas para ser eficaz como consultor, as quais vão muito além daquelas de um cientista de dados.

Dependendo do tipo de cliente que você consegue como consultor, seu trabalho provavelmente será metade ciência de dados e metade para manter a empresa andando. O trabalho terá certo fluxo: você trabalhará em um projeto para uma empresa-cliente e, à medida que o projeto se aproxima do fim, você ficará perto de fechar a venda de outro projeto para uma empresa diferente. Além disso, o trabalho terá um ritmo caótico: em um mês, haverá três empresas querendo seu tempo e, no mês seguinte, ninguém passa trabalho.

Captação de clientes é muitas vezes a parte a mais difícil de ser um consultor independente, pois exige dedicação e uma boa rede de contatos. A maioria dos clientes tende a vir por recomendações de antigos colegas ou clientes. Quanto mais pessoas conhecem um consultor e o recomendam, mais trabalho entrará. Assim, para que um consultor seja bem-sucedido, muitas pessoas (preferencialmente, aquelas com autoridade para contratar um consultor) têm de saber sobre a especialidade do consultor. Quanto mais diversificada for a rede de contatos entre as diferentes empresas e setores,

mais provável é que o trabalho entre em momentos diferentes. Ter uma boa rede de contatos assim requer que o consultor tenha trabalhado com muitas empresas diferentes antes, seja tendo trocado muito de trabalho antes na carreira dele ou tendo sido consultor em uma empresa maior.

Se for bem-sucedido como um consultor independente, terá a oportunidade de começar a contratar mais pessoas e fazer o negócio crescer. A empresa pode começar só com você e aumentar para uma equipe de cinco até chegar em 100. Como CEO e fundador da empresa, você poderá liderá-la na direção desejada com a cultura desejada. Ficando com parte do dinheiro que todos os outros consultores trazem para a empresa, você pode ficar rico. Embora esse resultado seja raro, ser um consultor independente pode ser de longe o mais rentável de todos os caminhos tratados neste capítulo.

### **16.3.1 Benefícios da consultoria independente**

Como consultor independente, você começa a ser seu próprio chefe, o que significa que começa a escolher se aceita ou não possíveis projetos de ciência de dados, que abordagem seguir e como apresentar os resultados. Não precisa depender de mais ninguém, o que para algumas pessoas pode ser incrivelmente libertador. Se puder manter suas despesas sob controle, o valor elevado e uma carteira consistente de clientes, sua empresa poderia ser potencialmente lucrativa. Você tem a oportunidade de ganhar o dobro do que se trabalhar em uma empresa. Se quiser trabalhar de casa ou tirar uma folga, você pode fazer isso sem ter que argumentar com alguém.

Você tem a propriedade daquilo que faz também. Se criar um método interessante para resolver um problema, pode decidir patentear-lo ou introduzi-lo no mercado como produto da sua empresa. Ninguém pode tirar o trabalho de você, ao passo que se trabalhar para a empresa de outra pessoa, essa pode reivindicar suas ideias como propriedade intelectual. Se puder criar um portfólio de produtos úteis, esse portfólio pode sustentá-lo por anos.

Prestar consultoria pode ser divertido! Há certo fator de emoção viajar pelo país, ajudar as pessoas com suas ideias e fazer tudo isso com o nome da sua própria empresa. Pode ser uma boa validação ter muitas pessoas que



querem pagar pelo seu tempo. Também pode ser ótimo apresentar uma solução que o cliente goste e saber que você é o autor do trabalho.

### 16.3.2 Desvantagens da consultoria independente

As desvantagens de ser um consultor independente são espantosas (tão difíceis, na verdade, que usaremos uma fonte em negrito nesta seção):

- **A consultoria independente é extremamente estressante.** Receber dinheiro em um mês depende inteiramente se decidirem contratá-lo, o que muitas vezes depende de fatores fora de seu controle (como orçamento da empresa). Por outro lado, é possível se encontrar com mais trabalho do que pode abraçar e precisa descobrir qual projeto deixará de lado. Muitas vezes, você precisa vender projetos de consultoria antes de conseguir acesso total aos dados para ver se o projeto seria até mesmo viável, e precisa descobrir o que fazer se não for. Há mil maneiras que ser um consultor pode mantê-lo acordado à noite.
- **A consultoria independente pode fazê-lo ficar sem dinheiro.** Caso se torne consultor independente em tempo integral e não conseguir encontrar oportunidades, perderá dinheiro rapidamente. Mesmo que possa assinar um contrato para dar consultoria, muitas vezes as grandes empresas não pagarão em até 90 ou 120 dias depois que o trabalho seja entregue, o que pode representar meio ano depois de ter começado. Se não conseguir lidar com esses altos e baixos nada saudáveis no fluxo de caixa, não poderá ser um consultor.
- **Não terá ninguém para recorrer.** Se estiver trabalhando sozinho como consultor, as pessoas chegaram em você porque têm um problema que não conseguem resolver, então, provavelmente, não haverá ninguém com quem discutir ideias. Você está sozinho. Se estiver tendo problemas para executar uma análise ou colocar um modelo em funcionamento, será forçado a encontrar uma solução sozinho; caso contrário, terá de dizer ao cliente que falhou.
- **Seu trabalho não será apenas ciência de dados.** O tempo que gastará fazendo marketing, negociando vendas, escrevendo contratos e cuidando

da contabilidade será imenso comparado com o tempo que passa fazendo ciência de dados. Ser um bom cientista de dados não é o suficiente; todo esse trabalho é necessário para sua empresa de consultoria sobreviver.

### **16.3.3 Como se tornar um cientista de dados líder**

Para tornar-se um consultor independente, você precisa dispor de um significativo conjunto de competências de ciência de dados e um histórico de resolução de problemas com independência no trabalho, assim como será muito importante dispor de uma rede de pessoas que conheçam suas competências, seja trabalhando em várias empresas ou (ainda melhor) em uma grande empresa de consultoria.

Essa consultoria poderá ser testada ao realizar trabalho freelance de ciência de dados em seu tempo livre. Construa um site, publique no LinkedIn e deixe que as pessoas saibam que está disponível para ajudar. Se puder conseguir clientes, aprenderá mais sobre consultoria enquanto faz trabalho freelance à noite. Se achar que não tem energia para fazer trabalho noturno, provavelmente não gostará de ser consultor. Se não conseguir clientes freelance, é sinal de que sua rede não é grande o suficiente, e você deve se concentrar nisso primeiro.

Se achar que tem tanto trabalho freelance que é difícil continuar com sua atividade de tempo integral, é sinal de que agora é um bom momento para fazer a transição para ser um consultor independente de tempo integral. A essa altura, é possível começar a focar bastante em consultoria e, se conseguir encontrar uma carteira dos principais clientes com a qual começar, é possível sair do seu emprego e mudar para a consultoria em tempo integral.

## **Sair da ciência de dados**

Uma última via é sair completamente da ciência de dados. Talvez você pense que o trabalho não é mais interessante. Talvez a carga de trabalho não se alinhe com suas necessidades de equilíbrio entre trabalho e vida pessoal. Talvez tenha se encontrado em um cargo no qual tenha de usar a ciência de dados para fins antiéticos, e você não consegue mais fazer isso. Há muitas razões pelas quais a ciência de dados não é a área certa para todos, e não há vergonha alguma nisso!

É difícil dar sugestões a respeito de como sair da área porque seu método depende muito da sua próxima área. Um currículo de ciência de dados pode gerar uma contratação fácil em áreas relacionadas, como desenvolvimento de software ou engenharia. Pode ser mais difícil fazer a transição para outras áreas. Porém, assim como é possível fazer com que funções anteriores soem parecidas com ciência de dados (Capítulo 6), é possível tentar fazer com que funções da ciência de dados sejam semelhantes com outra área.

Se realmente sair da área, talvez possa querer retornar mais tarde. Depois de deixar a ciência de dados, desde que faça um pequeno projeto periodicamente ou tente acompanhar a área, pode ser viável rever o que perdeu quando estiver pronto para retornar. Nesse caso, será como recomeçar desde o início este livro, apenas com mais experiência.

Embora a ciência de dados tenha muito marketing e agora esteja em alta como uma boa área, não se sinta pressionado em ter de ficar; sua felicidade é a coisa mais importante. Faça o que for certo para você!

## **16.4 Como escolher seu caminho**

Neste capítulo, apresentamos três caminhos para evoluir em ciência de dados, embora haja muitas outras opções, podendo parecer que o leque oferecido é muito grande e, em sua maioria, impossível tentar um caminho antes de assumir um significativo compromisso. Como saber qual caminho de carreira é o certo para você?

A verdade é que não tem como saber. Não se pode saber qual é a escolha “certa” porque inexiste uma escolha certa. Essas decisões dependem das empresas com as quais está trabalhando, das pessoas ao seu entorno e dos seus interesses pessoais nesse momento da sua vida. Você só pode fazer a escolha que sente ser a melhor para si e não se preocupar demais com as oportunidades perdidas.

Essa lição apareceu repetidas vezes neste livro. Assim como não existe uma forma correta de aprender competências da ciência de dados ou um tipo certo de empresa para trabalhar, não há uma forma melhor de navegar pelas áreas mais experientes da sua carreira. Somente pode fazer o que é melhor para si com o conhecimento que tem. Esperamos que este capítulo tenha proporcionado conhecimentos suficientes para facilitar um pouco a navegação por essas opções de carreira.

## **16.5 Entrevista com Angela Bassa, chefe da ciência de dados, engenheira de dados e machine learning na iRobot**

Angela Bassa, bacharel em matemática, é diretora da iRobot, onde supervisiona o trabalho de engenharia de dados, ciência de dados e machine learning de toda a empresa. Anteriormente, trabalhou como consultora, gerente sênior e diretora de análises.

## **Como é o dia a dia de gerente?**

Depende muito da complexidade da empresa, que muitas vezes depende também do tamanho dela. Quando há três pessoas, há pontos de conexões entre elas; quando há sete pessoas, você precisa de mais pontas para conectar todos uns aos outros. Se for necessário coordenar diferentes produtos, equipes, objetivos e cronogramas, então demanda muitas reuniões. Passo cerca de um terço do meu dia nessa coordenação estratégica para garantir que estamos trabalhando nas coisas certas, da forma certa e pela razão certa. Outro terço passo trabalhando com minha equipe, geralmente aconselhando e os ajudando com contexto ou feedback. O terço final é administrativo. Por exemplo, o orçamento está alinhado? Todos têm dinheiro para o treinamento e o desenvolvimento para os quais se inscreveram? Se houver uma conferência de mulheres muito interessante se aproximando e houver várias vagas abertas na minha equipe, será que quero patrocinar essa conferência?

## **Quais são os sinais de que deve subir de nível sendo um colaborador individual?**

Decidir tornar-se um gerente demanda muita introspecção, autoconsciência e mente aberta. Encontrar o ponto em que se tornar gerente representa maior probabilidade de sucesso tem muito a ver com estar em um momento (profissional e pessoalmente) para fazer uma transição tão grande. A gerência é uma profissão diferente: tem um conjunto de competências diferentes e um perfil de risco distinto. Se você não fizer um bom trabalho como colaborador individual, seu destino é por sua conta e risco. Quando gerencia outras pessoas, sua responsabilidade envolve o acesso de outra pessoa a planos de saúde ou a capacidade de outra pessoa pagar o aluguel. Acho que qualquer pessoa pode ser gerente e, se estiver preocupado em virar gerente, é provável que seja um excelente candidato.

## **No fim, é preciso fazer essa transição de sair da função de colaborador individual?**

A ciência de dados como profissão é tão nova que ainda existe uma grande

autosseleção de quem escolhe ingressar nessa profissão. Muitos de nós somos aquelas pessoas ambiciosas que vão lá e fazem acontecer: estamos abrindo o caminho de uma carreira porque somos o tipo de pessoas que fazemos isso com nossas vidas. Mas se olharmos para outros caminhos de carreira, como a contabilidade, eles não estão indo de um lado para o outro; é totalmente possível ser um contador sênior durante muito tempo. O que pode acontecer é chegar em um limite de crescimento de remuneração ou de conhecimento. Se for algo que atenda aos seus objetivos de carreira, não vejo problema algum uma pessoa gostar de estar exatamente onde se encontra. Ainda assim, há muitos profissionais talentosos que chegam aos montes à ciência de dados vendo-a como uma carreira, mas, se eu encontrar alguém com vontade e astúcia, é mais provável que terei uma tendência maior a esse tipo de personalidade ao decidir quem contratar.

### **Que conselhos você tem para alguém que quer ser uma liderança técnica, mas não está exatamente pronto para isso?**

Encontre uma pessoa que seja um porto seguro para ter uma conversa franca e aberta, de modo que não fique guardando seus pensamentos só para você. Isso pode ajudá-lo a receber feedback concreto e a compreender no que precisa melhorar. Ter alguém que já tenha tido êxito em uma função de liderança técnica também pode ser útil para ajudá-lo a entender o que eles precisaram para conseguir e ter sucesso nessa função. Verá seus pontos cegos quando investigar seu próprio conjunto de competências. É engraçado: aceitamos que é preciso de muita comunicação e colaboração para crescer e, como no ditado, de que é preciso uma aldeia para criar uma criança; mas profissionalmente, esperamos que todos sejam capazes de fazer tudo por conta própria. A melhor maneira de crescer é encontrar pessoas que o apoiem e estar disposto a ouvir o feedback delas.

### **Qual é seu conselho para cientistas de dados iniciantes e aspirantes?**

Meu primeiro conselho é ser *humilde*. É fácil acreditar que somos reis e rainhas, já que a ciência de dados é a profissão “mais sexy” e que qualquer empregador deve jogar pétalas de rosas aos nossos pés quando

caminhamos. É tão importante lembrar-se de que são necessárias muitas pessoas para que um produto seja bem-sucedido e que, só porque a ciência de dados está com a atenção voltada para ela, não significa que seja melhor ou especial.

O segundo conselho é ser *gentil*. É fácil criticar a nós mesmos, especialmente porque a ciência de dados é tão ampla e pode significar tantas coisas. Se for bom em análise, mas não tanto em engenharia de machine learning, talvez sinta que você não é um cientista de dados “de verdade”. Mas você é! Há muitas maneiras de brilhar!

## Resumo

- A gerência é um excelente caminho para quem quer ajudar os outros, além de estarem dispostas a desistir de fazer ciência de dados. Seguir esse caminho pode, no fim, fazê-lo subir muitos degraus na empresa.
- Um cientista de dados líder consegue liderar a parte técnica e é responsável por outras pessoas. Essa função é uma opção excelente para seguir na parte técnica enquanto auxilia os outros.
- A consultoria independente é muito estressante e arriscada, mas também é possivelmente recompensadora. Será preciso ter uma rede forte de conexões para encontrar trabalho de forma consistente.

## Recursos dos capítulos 13–16

### Livros

*The Design of Everyday Things*, de Don Norman (Basic Books)

Esse livro clássico apresenta ideias da área do design e discute como pensar sobre design em qualquer tipo de trabalho. Ser capaz de entender um usuário e como o design influencia o que ele faz é fundamental para o sucesso de um produto. Ao ler este livro, você entenderá melhor as necessidades de seus stakeholders e diminuirá a chance de que seu projeto de ciência de dados falhe porque não é aquilo que o cliente quer.

*Self-Compassion: The Proven Power of Being Kind to Yourself*, de Kristin Neff, PhD (HarperCollins Publishers)



Se a parte do Capítulo 13 que discutiui como as pessoas se crucificam mentalmente quando falham soou familiar, talvez queira ler esse livro. Ele se aprofunda na batalha da autora com a autocrítica e sua trajetória para ser compreensiva com ela mesma. Essa trajetória é a que muitos cientistas de dados devem seguir, e esse livro é um grande guia para isso.

*Demystifying Public Speaking*, de Lara Hogan (A Book Apart)

Lara Hogan é uma reconhecida palestrante pública sobre liderança de engenharia e que escreveu esse livro para dar conselhos práticos e ajudar as pessoas a começarem a palestrar. Nessa publicação curta e envolvente, aborda táticas para tudo, desde escolher um tópico a fazer na apresentação até lidar com os nervos.

*R Packages*, 2ª ed, de Jennifer Bryan e Hadley Wickham (O'Reilly Media)

Este livro aborda os detalhes de como criar um pacote R, uma excelente forma de melhorar seu próprio fluxo de trabalho e como retribuir à comunidade. Em janeiro de 2020, o livro ainda estava em desenvolvimento, mas você pode encontrar a cópia do trabalho em andamento em <https://r-pkgs.org>.

*Resilient Management*, de Lara Hogan (A Book Apart)

Esse livro é um guia ótimo e curto para todo novo gerente, oferecendo conselhos e modelos para começar a conhecer seus colegas de equipe, fazer mentoria, ajustar expectativas e tratar de desafios. Mesmo que não esteja planejando ser gerente de pessoas, a publicação será muito útil se estiver começando a liderar projetos ou se tiver tido dificuldades em se comunicar com seus colegas de equipe.

*The Manager's Path: A Guide for Tech Leaders Navigating Growth and Change*, de Camille Fournier (O'Reilly Media)

Se estiver pensando em mudar da função de cientista de dados para ser gerente ou líder técnico, esse livro é perfeitamente direcionado a você. Escrito por Camille Fournier, antiga CTO da Rent the Runway, o livro trata de como pensar o trabalho de um gerente e deixar a ideia de ser um colaborador individual. Ler essa publicação o ajudará a compreender ideias que podem levar anos para se dar conta por si mesmo.

*The E-Myth Revisited*, de Michael E. Gerber (HarperCollins Publishers)

Embora esteja totalmente não relacionado à ciência de dados, esse livro é um ótimo recurso para pessoas que estão considerando atuar como consultores independentes ou iniciando seus próprios negócios de outra forma. Ele discute como a estrutura de pensar sobre seu trabalho necessita mudar enquanto dirige uma empresa. Você precisa transformar-se de uma pessoa focada em completar tarefas em alguém que precisa sistematizar tudo o que faz para manter a empresa sempre funcionando.

*High Output Management*, de Andrew S. Grove (Vintage)

Gerenciar bem uma empresa é uma empreitada complexa que requer muito pensamento estratégico. *High Output Management* divide os conceitos, utilizando casos de fácil compreensão, como a entrega de café da manhã, e constitui-se em um recurso útil para pessoas que pretendem compreender mais sobre a gerência de uma empresa. O livro foi lançado em 1983, mas tem sido atualizado e permanece atual.

## **Blogs**

“Making peace with personal branding”, de Rachel Thomas

<https://www.fast.ai/2017/12/18/personal-brand>

Rachel Thomas faz um excelente trabalho ao colocar sua marca na área da ciência de dados e nas redes sociais, e este post do blog oferece uma perspectiva sobre como fazê-lo ao mesmo tempo em que se sente confortável consigo mesmo.

Blog da Lara Hogan



<https://larahogan.me/blog>

Recomendamos dois dos livros de Lara Hogan na seção anterior, e o blog dessa autora está igualmente cheio de bons conselhos sobre as competências pessoais de que necessita para ter sucesso. Embora muitas das suas publicações se concentrem em gerentes, também dá conselhos que são aplicáveis a qualquer pessoa, incluindo o que fazer se seu gerente não apoiá-lo, como dar feedback e lidar com suas emoções quando pensa “por que a liderança não é só...?”.

“The art of slide design”, de Melinda Seckington

<https://missgeeky.com/2017/08/04/the-art-of-slide-design>

Essa série de cinco partes (cada publicação subsequente tem um link na parte inferior) é uma aula sobre criar apresentações eficazes. Seckington compartilha os princípios do design de slides – maximize o sinal, minimize o ruído, faça com que as informações importantes se destaquem, conte uma história e seja consistente – e os ilustra com muitos exemplos e contraexemplos.

“Overcoming social anxiety to attend user groups”, de Steph Locke

<https://itsalocke.com/blog/overcoming-social-anxiety-to-attend-user-groups>

Se a ansiedade social o tiver impedido de participar de meetups ou conferências, leia esse post de Steph Locke. Ela aborda preocupações comuns como: “Não conheço ninguém” ou “Como falo com as pessoas?” com conselhos breves e práticos.

“How to ask for a promotion”, de Rebecca Knight

<https://hbr.org/2018/01/how-to-ask-for-a-promotion>

Esse artigo compartilha dicas de dois coaches de liderança sobre como pedir uma promoção. Cada dica, como “plante a semente” e “faça uma pesquisa”, vem com um parágrafo de exemplos específicos.

# Epílogo

Bem, temos certeza de que tratamos de muitos assuntos neste livro. Começamos definindo ciência de dados e as competências necessárias, depois vimos como se preparar e conseguir um emprego de ciência de dados e discutimos como crescer na área. Nesses 16 capítulos conhecemos diferentes tipos de empresas, fizemos testes de unidade para modelos de produção e vimos como se tornar um gerente.

Ao olhar o conjunto do livro, algumas tendências parecem fluir por ele. Essas lições aplicam-se de diferentes formas a todos os momentos da trajetória de um cientista de dados. Para nós, esses três ideais fizeram com que nossas carreiras continuassem indo em frente:

- *Um cientista de dados precisa conseguir se comunicar.* Diversas vezes, as pessoas que entrevistamos para o livro mencionaram que o sucesso veio de comunicar o trabalho com eficácia. Independentemente da comunicação eficaz ser a elaboração de um relatório para um executivo, colaborar com uma equipe de engenharia em um modelo ou ser capaz de falar de uma forma que aquelas pessoas que não são cientistas de dados compreendam, pode ajudá-lo no processo de encontrar um emprego e trabalhar com outras pessoas nesse cargo.
- *Um cientista de dados precisa ser proativo.* É extremamente raro que um cientista de dados receba um problema perfeitamente bem formado e as ferramentas para solucioná-lo. Os cientistas de dados precisam tentar encontrar dados de forma proativa, criar ideias para modelos e fazer experimentos. Ser proativo e fazer algo como um portfólio, por exemplo, irá também ajudá-lo a conseguir um emprego. Quanto mais você tomar iniciativa e encontrar soluções para os problemas, melhor.
- *Um cientista de dados precisa de uma comunidade.* Ninguém segue em uma carreira sem a ajuda de outros, mas, como parte de uma área nova e em rápido crescimento, cientistas de dados beneficiam-se especialmente de boas relações profissionais. Esses relacionamentos podem ocorrer de

muitas formas. Alguém pode indicá-lo como palestrante em uma pequena reunião, o que dois anos mais tarde pode levá-lo a palestrar em uma conferência internacional. Um mentor pode revisar seu currículo e encaminhá-lo para uma vaga na empresa dele. Um gerente pode ajudá-lo a preencher a lacuna com stakeholders e sugerir áreas para crescimento pessoal. Ou um colega pode simplesmente ser empático e animá-lo após um dia difícil no trabalho. Vale a pena investir tempo na construção dessas relações para enfrentar os muitos desafios da carreira.

Esperamos que tenha gostado de ler o livro; sem dúvida, gostamos de escrevê-lo. Ao criar este livro, descobrimos que muitas de nossas próprias experiências foram colocadas nele. Várias vezes, durante o processo de redação, nós até mesmo relemos alguns trechos para repensar nossas próprias decisões na carreira. Desejamos o melhor a você em sua carreira em ciência de dados!

# APÊNDICE

## Perguntas da entrevista

Muitas vezes, ao preparar-se para uma entrevista, o mais útil é imaginar como ela será. Sentir-se confortável em responder a perguntas e a pensar de uma maneira adequada ao ritmo rápido de uma entrevista pode ser a diferença entre conseguir o emprego ou não. Por isso, trazemos exemplos de perguntas de entrevista para se pensar e entender. Você deve ver essas perguntas em conjunto com o Capítulo 7, o qual discute como entender o processo de entrevista na sua totalidade.

As perguntas deste Apêndice são classificadas em cinco categorias:

- Programação e desenvolvimento de software
- SQL e bancos de dados
- Estatística e machine learning
- Comportamental
- Problemas de lógica

Trata-se uma vasta gama de tópicos, sendo impossível estudar para as milhares de questões que podem ser feitas. Uma empresa pode pedir que você inverta uma árvore binária enquanto outra faz apenas perguntas comportamentais e sobre Python. É por isso que recomendamos que, antes das entrevistas na empresa, você pergunte que tipos de perguntas esperar. Não dirão as perguntas exatas, é claro, mas o gerente de contratação ou recrutador pode dar uma ideia geral para que você consiga focar sua preparação. Eles podem dizer, por exemplo: “Em sua primeira entrevista, você responderá a algumas perguntas de SQL em um quadro branco. Depois, terá duas entrevistas comportamentais seguidas, sendo uma com um engenheiro e outra com um cientista de dados. Por fim, um dos nossos engenheiros de machine learning fará perguntas sobre seus projetos anteriores de ciência de dados”.

É extremamente improvável que você veja somente as perguntas abordadas neste Apêndice durante o processo de busca de emprego. É por isso que não apenas respondemos cada pergunta (com o texto que diríamos em voz alta e o código que escreveríamos no quadro branco), mas também fazemos observações sobre o que pensamos ser uma resposta correta. Todas as respostas são dadas em primeira pessoa, do ponto de vista de um cientista de dados hipotético que tem uma combinação de experiências semelhantes às das autoras deste livro. Para algumas das perguntas, baseamo-nos nas experiências que tivemos em empregos anteriores; você deve tentar apresentar seus próprios exemplos para essas questões.

Algumas das perguntas originam-se de nossas experiências compartilhadas passando por muitas entrevistas; outras basearam-se em colegas cientistas de dados. Agradecemos a todos que colaboraram para tornar este Apêndice muito mais útil!

## A.1 Programação e desenvolvimento de software

### A.1.1 FizzBuzz

*Escreva um programa que imprima os números de 1 a 100. Mas, para múltiplos de 3, imprima "Fizz" em vez do número e, para os múltiplos de 5, imprima "Buzz". Para números múltiplos de 3 e 5, imprima "FizzBuzz".*

### Resposta de exemplo

A seguir, um pseudocódigo para uma solução do problema:

```
for (i in 1 to 100) {  
  if (i mod 15) {  
    print("FizzBuzz")  
  } else if (i mod 5) {  
    print("Buzz")  
  } else if (i mod 3) {  
    print("Fizz")  
  } else {  
    print(i)  
  }  
}
```



O programa se repete para os números de 1 a 100. Para cada repetição, ele verifica primeiro se o número é divisível por 15 e, em caso afirmativo, imprime “FizzBuzz”. Caso contrário, ele verifica se o número é divisível por 5 e, em caso afirmativo, imprime “Buzz”. Caso contrário, ele verifica se o número é divisível por 3 e imprime “Fizz”; se for, e se nenhum deles for verdadeiro, o número é impresso.

## Observações

Esse problema é uma pergunta de entrevista extremamente conhecida no desenvolvimento de software, tendo sido criada por Imran Ghory e popularizada por Jeff Atwood (<https://blog.codinghorror.com/why-cant-programmers-program>), razão pela qual é comum receber essa pergunta exata como parte de uma entrevista de ciência de dados. Suas duas tarefas principais são como repetir sobre o conjunto de todos os números (no exemplo, usamos um for para o loop) e como verificar o que deve ser impresso em cada número. Um erro comum é verificar se o número é divisível por 3 ou 5 antes de verificar se é divisível por 15, mas qualquer número divisível por 15 também é divisível por 3 ou 5. Assim, se 3 ou 5 forem verificados primeiro, “Fizz” ou “Buzz” poderão ser impressos nos casos em que “FizzBuzz” deveria ser.

Embora a solução que propusemos seja simples, há formas de melhorá-la. Em algumas linguagens, incluindo R e Python, você pode usar uma abordagem de programação funcional mais limpa, usando purrr em R ou compreensão de lista em Python. Também é possível criar uma função generalizada que, como a inserção pega a lista de múltiplos para verificar e imprime palavras nesses múltiplos, resulta em uma lista. Dependendo de como a entrevista estiver se desenvolvendo, talvez queira falar sobre as maneiras que gostaria de melhorar a resposta.

Por diversão, confira o FizzBuzz Enterprise Edition (<https://github.com/Enterprise-QualityCoding/FizzBuzzEnterpriseEdition>) ou um modelo de machine learning FizzBuzz no TensorFlow (<https://joelgrus.com/2016/05/23/fizz-buzz-in-tensorflow>).

### A.1.2 Informe se um número é primo

*Escreva uma função que, dado um número, retorna true se for um número primo e false, caso contrário. Suponha que não haja uma função incorporada para verificar se um número é primo.*

## Resposta de exemplo

A seguir, um pseudocódigo para uma solução do problema:

```
is_prime = function(n){  
  for (i in 2 to n / 2) {  
    if ((n mod i) == 0) {  
      return FALSE  
    }  
  }  
  return TRUE  
}
```

Um *número primo* é aquele somente divisível por 1 e por ele mesmo. O programa repete todos os números a partir de 2 até a metade do número dado e verifica se o número indicado é divisível por eles. Se for, a função retorna false e para. Se passar por todo o loop for sem parar, a função retorna true.

## Observações

Da mesma forma que o FizzBuzz, esse problema testa se você consegue escrever um loop for e uma função. Também precisa saber como parar de repetir quando uma condição é alcançada, a fim de que possa retornar com segurança true no fim se o loop for se completar. Dá para adicionar pequenos truques, como não precisar verificar se o número é divisível por todos os números menores que ele ou pela metade dele – apenas aqueles menores que a sua raiz quadrada. Mas o ponto principal do problema é simplesmente testar se você consegue escrever uma função que funcione.

### A.1.3 Como trabalhar com Git

*Você pode falar sobre um momento em que usou o Git para colaborar em um projeto?*

– Alex Hayes

## Resposta de exemplo

Em meu último emprego, criei um pacote R, `funneljoin`, com dois colegas durante uma tarde de hackathon, usando Git desde o início. Passamos a primeira hora programando em pares em um computador e, depois, fizemos uma lista de tarefas para dividir entre nós. Cada um criou um ramo diferente para trabalhar em nossas tarefas, o que nos permitiu facilmente mesclá-las novamente no fim. Usar o Git garantiu que nunca sobrescrevemos acidentalmente o trabalho de outra pessoa. Ao comprometer-se cedo e continuamente à medida que progredimos, sabíamos que seria possível sempre regressar caso decidíssemos que uma forma anterior de implementar uma funcionalidade era melhor. Por fim, usar o GitHub fez com que, mais tarde, qualquer pessoa da empresa pudesse baixar o pacote e começar a usá-lo imediatamente.

Desde essa tarde, permaneci como mantenedor do pacote. Continuo usando recursos do Git como ramificações para poder criar protótipos de recursos sem mesclá-los até que tenham sido completamente testados.

## **Observações**

Se fizeram essa pergunta e você não tiver colaborado usando o Git em um projeto antes, poderá falar sobre como usou o Git para um projeto pessoal. Essa pergunta está testando se usou o Git, podendo explicar que utilizou recursos diferentes (como ramificações ou bifurcações), mesmo que só para você. É possível acrescentar como você adaptaria suas práticas se estivesse colaborando com outras pessoas, usando mais ramificações, por exemplo, ou aderindo a uma estrutura consistente de mensagens de confirmação.

Se já não tiver usado o Git, seja honesto sobre isso na entrevista. Recomendamos que tente aprender isso antes de ter de fazer muitas entrevistas.

### **A.1.4 Decisões de tecnologia**

*Dada uma tela totalmente em branco, como você escolhe sua stack de tecnologia?*

– Heather Nolis

## **Resposta de exemplo**

Essa pergunta é interessante de responder porque realmente depende de ter o projeto em mãos. Minha decisão para um recurso de tecnologia baseia-se sobretudo em equilibrar o que seria o mais simples para implementar com aquilo que seria o mais fácil com que todos os demais trabalhem. Deixe-me dar dois exemplos de como escolhi os recursos de tecnologia e o que aprendi com eles.

Em um projeto anterior, tive de desenvolver um produto novo inteiramente a partir do zero, e eu era a única cientista de dados em minha equipe. Escolhi usar .NET e F# porque estava bastante familiarizado com eles, e, por conta disso, conseguimos rapidamente um produto que funcionasse. A desvantagem foi que, pelo fato de a linguagem F# ser tão incomum, quando chegou a hora de contratar um cientista de dados para assumir o cargo não conseguimos encontrar ninguém que já tivesse o conhecimento necessário. Em retrospectiva, a utilização de .NET e F# não foi a decisão certa.

Em um projeto mais recente, fui encarregada de criar uma API de machine learning. Estava no meio de uma equipe de engenharia que trabalhou com microsserviços, então decidi criar uma API REST em linguagem R como um contêiner Docker. Embora nunca tivesse usado contêineres Docker, fiz essa escolha porque sabia que seria mais fácil para a equipe mantê-la. A partir desse projeto, aprendi muito sobre Docker e contêineres, e o trabalho que criei foi capaz de integrar bem.

## **Observações**

Ao entrevistar candidatos na T-Mobile, Heather Nolis geralmente faz uma pessoa escolher um projeto e descrever as decisões tomadas, as decisões que outras pessoas tomaram e como fariam as coisas de forma diferente, sabendo o que sabem agora. Seu entrevistador talvez não lhe pergunte diretamente essas coisas, mas, ainda assim, todas valem a pena de serem incluídas.

Seja qual for a resposta dada, inclua muitas referências às decisões que teve de tomar no trabalho que já realizou (que podem incluir projetos paralelos ou trabalhos de formação). O objetivo dessa questão é ver quanto raciocínio você desenvolve em escolher a tecnologia certa para o projeto

certo. Ter escolhido recursos de tecnologia que acabaram sendo problemáticos é bom, desde que tenha aprendido com o problema. Na verdade, ter aprendido com as coisas é ainda melhor do que fazer tudo certo de primeira, porque mostra que você pode mudar.

### **A.1.5 Pacote/biblioteca frequentemente utilizado**

*Qual pacote R ou biblioteca Python você usa com frequência e por quê?*

#### **Resposta de exemplo**

Não é um pacote, mas realmente gosto muito do conjunto de pacotes que compõem o tidyverse em R. Os pacotes fazem de tudo, desde leitura de dados a limpeza até transformação, visualização e modelagem.

Gosto especialmente de trabalhar no dplyr porque, graças ao pacote conectado dbplyr, posso escrever o mesmo código se estou trabalhando com uma tabela local ou remota, pois o dbplyr traduz o código dplyr para SQL. Usar dbplyr no meu último trabalho significou que eu poderia ficar no RStudio para todo o meu fluxo de trabalho, mesmo que todos os nossos dados fossem armazenados no Amazon Redshift e precisassem de consultas SQL para acessar. Eu usava o dbplyr para fazer resumos e filtragem e, em seguida, pegava os dados localmente se precisasse fazer operações ou visualizações mais difíceis.

No geral, realmente gosto da filosofia de Hadley Wickham, um desenvolvedor de tidyverse importante: que o gargalo ao programar muitas vezes é pensar no tempo, não no tempo computacional, e que você deve construir ferramentas que funcionem perfeitamente juntas e permitem que você traduza seus pensamentos em código com rapidez.

#### **Observações**

O entrevistador não está procurando uma resposta específica nesse caso. Em vez disso, quer saber se você (a) programa suficientemente em qualquer linguagem para ter um pacote usado com frequência e (b) pode explicar como e por que usa esse pacote. Essa resposta também dá ao entrevistador uma ideia de que tipo de trabalho você faz no dia a dia. Não se esqueça de explicar o que o pacote faz, em especial se for um pacote de um nicho. Se

outro pacote for mais amplamente usado para a tarefa, explique seu raciocínio subjacente à escolha desse pacote, pois apresenta uma ideia mais ampla sobre as alternativas existentes. Por fim, não se preocupe em escolher a biblioteca mais “avançada”, como, por exemplo, uma biblioteca de deep learning, para impressionar o entrevistador. Em teoria, essa pergunta é uma das mais fáceis e (potencialmente) mais divertidas de responder, por isso, não pense demais na resposta.

### **A.1.6 R Notebooks do Jupyter ou Markdown**

*O que é um arquivo R Markdown ou Notebook Jupyter? Por que você usaria um arquivo R Markdown ou um Notebook Jupyter em vez de um script em R ou Python? Quando um script é melhor?*

#### **Resposta de exemplo**

Vou responder no caso de R e R Markdown, mas a ideia-base é a mesma para Notebooks Jupyter e Python. Arquivos R Markdown são maneiras de escrever código R que permitem colocar texto e formatação em torno do código. Em certo sentido, mesclam o código e os resultados de uma análise com a narração e as ideias da análise. Usando um arquivo R Markdown, é possível ter uma análise que seja mais fácil de reproduzir do que o código R bruto, com documentação separada sobre o que foi a análise. Em teoria, seu R Markdown seria formatado de forma tão clara que, ao renderizar o arquivo de saída, você poderia apresentar a um stakeholder o resultado como HTML, documento do Word ou PDF.

Os arquivos R Markdown são ótimos para análises reproduzíveis, mas menos úteis quando está escrevendo o código que implantará ou usará em outros locais. Digamos que você tem uma lista de funções que deseja usar em vários outros locais (como um para carregar os dados de um arquivo). Pode fazer sentido escrever um script R que crie todas as funções e mantenha o script separado de uma análise individual. Ou, se você quisesse usar R com o pacote Plumber para criar uma API da web, não seria necessário um arquivo R Markdown para isso.

#### **Observações**

Essa questão é uma verificação feita pelo entrevistador para ver se você tem experiência em fazer uma análise reproduzível. Muitas pessoas que usam R ou Python usam scripts de maneira objetiva e não pensam sobre como compartilhar os resultados com outras pessoas. Ao mostrar que compreende o objetivo do R Markdown e dos Notebooks Jupyter, você mostra que está pensando em como tornar seu código mais utilizável. Se não tiver usado arquivos R Markdown ou Notebooks Jupyter, definitivamente teste um deles.

Não se preocupe em compreender as versões R e Python; qualquer uma delas funciona.

### **A.1.7 Quando você deve escrever funções ou pacotes/bibliotecas?**

*Em que ponto você deve transformar seu código em uma função?  
Quando deve transformá-lo em um pacote ou biblioteca?*

#### **Resposta de exemplo**

Em geral, se eu perceber alguma vez que estou copiando e colando código, provavelmente é um sinal de que deveria fazer dele uma função. Se precisar executar código em três conjuntos de dados diferentes, por exemplo, eu deveria fazer uma função e aplicá-la a cada um em vez de copiar o código três vezes. A biblioteca `purrr` em R ou a compreensão de lista em Python facilitam muitas vezes a aplicação de uma função.

Descobri que os pacotes e as bibliotecas são melhores quando você tem código que abrange vários projetos distintos na equipe. No meu atual trabalho, temos muitos dados que armazenamos em S3, mas queremos analisar localmente. Em vez de copiar e colar funções para acessar o código em cada projeto, criei uma biblioteca que poderia ser chamada de todos eles. A desvantagem das bibliotecas é que se você as mudar, é preciso mudar todos os projetos que usam a biblioteca, mas, para as funções principais, essa abordagem muitas vezes vale a pena.

#### **Observações**

A questão é um tanto fácil porque tem uma resposta certa: “Tanto quanto

possível”. Geralmente, copiar e colar código repetidamente é uma prática ruim; um cientista de dados deve fazer funções de forma que o código seja mais fácil de ler e entender. Por essa razão, tanto quanto possível, adicione exemplos que mostrem que você entende o valor de reutilizar código como funções ou pacotes. Já fez uma função e a reutilizou muito? E uma biblioteca? Fale sobre essas situações o máximo que puder.

### A.1.8 Exemplo de manipulação de dados em R/Python

Aqui está uma tabela chamada tuítes. Os dados têm a conta que enviou o tuíte, o texto, o número de curtidas e a data de envio. Escreva um script para obter uma tabela com uma linha por pessoa com uma coluna que seja o número mínimo de curtidas, chamada de min\_curt, e uma coluna do número total de tuítes, denominada nb\_tuítes. Deve ser apenas para tuítes enviados após 1º de setembro de 2019. Você também precisa eliminar todas as duplicatas na tabela primeiro.

account_name (nome_conta)	Text (texto)	nb_likes (nb_curtidas)	date (data)
@vboykis	Ciência de dados é...	50	01/10/2019
@Randy_Au	É difícil quando...	23	01/05/2019
@rchang	Algumas notícias...	35	01/01/2019
@vboykis	Minha newsletter...	42	23/11/2019
@drob	Meu melhor conselho...	62	01/11/2019
...	...	...	...

### Resposta de exemplo em R

```
tweets %>%  
  filter(date > "2019-09-01") %>%  
  distinct() %>%  
  group_by(account_name) %>%  
  summarize(nb_tweets = n(), min_likes = min(nb_likes))
```

### Resposta de exemplo em Python

```
tweets = tweets[tweets.date > "2019-09-01"].  
drop_duplicates().  
groupby("account_name")  
  
tweets['nb_likes'].agg(nb_tweets="count", min_likes="min")
```



## Observações

Esse tipo de questão é uma mistura entre as de FizzBuzz e número primo (saber como fazer algo em R/Python) e SQL (analisando dados). Essa questão específica deve ser relativamente fácil para alguém que tenha feito análise de dados antes, mas pode enfrentar uma que tenha pegadinhas (como a necessidade de converter uma coluna de caracteres em uma coluna de data ou alterá-la de formato longo para largo), que pode não se lembrar no momento. Se você não se lembra de como fazer alguma coisa, basta dizer: “Não me lembro da sintaxe exata para  $X$ , então vou colocar algum pseudocódigo lá agora como um marcador de posição” e seguir em frente. Não precisa passar tempo demais travado em uma parte. Se a questão incluir algo mais incomum, é provável que o entrevistador considere que essa parte é um bônus em vez de um requisito para passar para a fase seguinte.

## A.2 SQL e bancos de dados

### A.2.1 Tipos de joins

*Explique a diferença entre um join à esquerda (left join) e um join interior (inner join).*

– Ludamila Janda e Ayanthi G.

### Resposta de exemplo

Os joins são formas de combinar dados de duas tabelas diferentes – uma tabela à esquerda e outra à direita – em uma terceira tabela. Joins funcionam conectando linhas entre as duas tabelas; um conjunto de colunas-chave é usado para localizar dados nas duas tabelas que são as mesmas e devem ser conectadas. No caso de um join à esquerda, cada linha da tabela à esquerda aparece na tabela resultante, mas as linhas da tabela à direita aparecem somente se os valores em suas colunas-chave aparecerem na tabela à esquerda. Em um join interno, entretanto, ambas as linhas da tabela esquerda e da tabela direita aparecem apenas se houver uma linha correspondente na outra tabela.

Na prática, você pode pensar em um join à esquerda como anexação de dados da tabela à esquerda, se ela existir (por exemplo, usar a tabela à direita como uma busca). Um join interno é mais como encontrar todos os dados compartilhados e fazer uma nova tabela somente dos pares.

## Observações

Janda gosta dessa questão como uma triagem inicial para cargos de começo de carreira, pois não é uma pergunta difícil e é um conhecimento importante para um candidato ter. Ela acha que se pode aprender muito sobre como o candidato escolhe responder. Há muitas respostas válidas, desde aquelas que são supercorretas, mas não são fáceis de compreender, até respostas muito simples de compreender, mas que perdem detalhes.

Observe que, em nossa resposta, não falamos de complexidades de linhas duplicadas que aparecem nos dados. Talvez valha a pena mencionar essas complexidades porque elas podem afetar os resultados, mas, mais provavelmente, elas são uma distração do objetivo que está tentando explicar.

### A.2.2 Carregando dados em SQL

*Quais são algumas maneiras diferentes de carregar dados em um banco de dados em primeiro lugar e quais são as vantagens e desvantagens de cada um?*

– Ayanthi G.

### Resposta de exemplo

Há muitas maneiras de carregar dados em um banco de dados, principalmente dependendo, em primeiro lugar, de onde os dados se encontram. Se os dados estiverem em um arquivo simples, como um arquivo CSV, muitas versões SQL têm programas para importar os dados. O SQL Server 2017, por exemplo, tem um assistente de importação e exportação. Essas ferramentas são fáceis de utilizar, mas não permitem uma grande personalização e não são reproduzíveis com facilidade. Se os dados forem provenientes de um ambiente diferente, como R ou Python, há drivers que permitem a transmissão dos dados para SQL. Um driver ODBC,

por exemplo, pode ser usado junto com o pacote DBI em R para mover dados de R para SQL. Esses métodos são mais reproduzíveis e programáticos para implementar, mas exigem que você obtenha os dados em R ou Python.

## Observações

Essa questão é realmente um teste para ver se você já carregou dados em um banco de dados. Se já o fez, não deverá ser muito difícil de descrever como o fez. Se ainda não tiver carregado dados em um banco de dados, isso poderá sinalizar ao entrevistador que você não tem experiência suficiente.

A parte da questão sobre vantagens e desvantagens quer verificar se você compreende que ferramentas diferentes são melhores em distintas situações. Às vezes, usar uma GUI para carregar dados é uma solução boa e fácil quando se tem um único arquivo. Em outras ocasiões, você vai querer configurar um script totalmente automatizado para carregar os dados continuamente. Quanto mais você puder mostrar que compreende as nuances do que e quando usar, melhor.

### A.2.3 Exemplo de consulta SQL

Aqui está a TABLE\_A de uma escola, contendo notas de 0 a 100 obtidas por alunos em várias turmas. Como calcularia a nota mais alta em cada turma?

Class (Turma)	Student (Aluno)	Grade (Nota)
Matemática	Nolis, Amber	100
Matemática	Berkowitz, Mike	90
Literatura	Liston, Amanda	97
Espanhol	Betancourt, Laura	93
Literatura	Robinson, Abby	93
...	...	...

## Resposta de exemplo

Aqui está uma consulta para encontrar a nota mais alta em cada turma:

```
SELECT CLASS, MAX(GRADE)
INTO TABLE_B
FROM TABLE_A
GROUP BY CLASS
```

Essa consulta agrupa os dados em cada classe e encontra a máxima a partir daí. Além disso, ela salva o resultado em uma nova tabela (TABLE\_B) para que os resultados possam ser consultados mais tarde.

## Observações

Essa pergunta é a mais simples sobre SQL; ela testa se você tem uma compreensão básica sobre agrupamento em SQL. As razões pelas quais as pessoas se confundem com essa questão incluem não ver o que agrupar (neste caso, a variável classe), ou acham que a questão é tão fácil que acabam complicando demais e não oferecem uma solução simples. Se estiver em uma entrevista e uma questão parecer ser fácil demais, pode ser que seja fácil mesmo.

Se essa solução não parecer óbvia para você, agora seria um bom momento para revisar como as variáveis de agrupamento funcionam no SQL.

Por fim, a linha INTO TABLE\_B era totalmente opcional, mas ela é uma boa base para a próxima questão.

### A.2.4 Continuação do exemplo de consulta SQL

*Considere a tabela da questão anterior. Se quiséssemos não só encontrar a nota mais alta em cada classe, mas também o aluno que obteve essa nota?*

## Resposta de exemplo

Supondo que tenhamos o resultado da questão anterior armazenada na TABLE\_B, podemos usá-la nesta solução:

```
SELECT a.CLASS, a.GRADE, a.STUDENT  
FROM TABLE_A a  
INNER JOIN TABLE_B b ON a.CLASS = b.CLASS AND a.GRADE = b.GRADE
```

Essa consulta seleciona todos os alunos e suas notas da TABLE\_A original, que têm classes com notas que aparecem na tabela de máximas, TABLE\_B. O join interno atua como um filtro para manter apenas as combinações de classe/nota que são as máximas, pois somente nesse caso a nota aparece na TABLE\_B. Alternativamente, poderíamos usar uma subconsulta para fazer a mesma coisa sem chamar a TABLE\_B:

```
SELECT a.CLASS, a.GRADE, a.STUDENT
TABLE_A
INNER JOIN (
  SELECT CLASS, MAX(GRADE)
  FROM TABLE_A GROUP BY CLASS) b
ON a.CLASS = b.CLASS AND a.GRADE = b.GRADE
```

## Observações

Embora esse problema tenha várias soluções, qualquer uma delas quase que com certeza requer mais do que uma única consulta da TABLE\_A e, por isso, essa questão pode facilmente confundir as pessoas. A solução pode parecer simples no papel, mas ser capaz de pensar nisso durante uma entrevista pode ser difícil. Se errar esse tipo de questão, ainda será possível passar na entrevista.

A solução não faz nenhuma previsão para o máximo. Na solução de exemplo, vários alunos seriam retornados. Pode valer a pena apontar esse fato ao entrevistador, porque mostra que você está prestando atenção aos casos extremos (edge cases).

## A.2.5 Tipos de dados

Quais desvantagens existem para armazenar uma coluna de datas como strings em um banco de dados? No SQL, por exemplo, e se armazenarmos uma coluna de datas como VARCHAR(MAX) em vez de DATE?

## Resposta de exemplo

Ter datas armazenadas como strings em vez de datas (por exemplo, armazenar 20 de março de 2019 como a string "20/03/2019") é uma situação comum em bancos de dados. Embora possa não perder informações, dependendo de como é feito, você pode experimentar problemas de desempenho. Primeiro, se os dados não forem armazenados como do tipo DATE, não seria possível usar a função MONTH(). Também não foi possível fazer coisas como encontrar as diferenças entre duas datas ou encontrar a data mínima na coluna.

Esse problema tende a acontecer muito quando você está carregando dados em um banco de dados ou limpando-o. Quanto mais cedo for possível

formatar corretamente os dados, mais fácil será a análise. É possível corrigir esses tipos de situações utilizando funções como CAST. Dito isso, se estiver carregando dados com centenas de colunas e houver muitos que nunca usará, pode não valer a pena o tempo de corrigir todos esses problemas.

## Observações

Ter dados armazenados em um tipo incorreto é um problema muito frequente. Isso não acontece apenas em bancos de dados; também pode acontecer em arquivos simples ou em tabelas em ambientes como R e Python. Essa questão quer verificar se você entende que, quando isso acontece, geralmente é ruim, e quando se deparar com essas situações, via de regra deve corrigi-las. Ser capaz de responder a questões como esta deve ser algo natural depois da experiência de limpeza de dados como parte de projetos de ciência de dados que começam com dados bagunçados.

## A.3 Estatística e machine learning

### A.3.1 Termos estatísticos

*Explique os termos média, mediana e modo a uma criança de 8 anos.*

– Allan Butler

### Resposta de exemplo

*Média, mediana e modo* são três tipos diferentes de médias. As médias nos permitem compreender algo sobre um conjunto inteiro de números com apenas um número que resume algo sobre todo o conjunto.

Suponha que fizemos uma pesquisa em sua turma para ver quantos irmãos cada pessoa tem. Você tem cinco pessoas em sua turma. Digamos que você acha que uma pessoa não tem irmãos, uma tem uma, uma tem duas e duas têm cinco.

O *modo* é o número mais comum de irmãos. Neste caso é 5, já que duas pessoas têm cinco irmãos em comparação com apenas uma pessoa que tem todos os outros números.

Para obter a média, você obtém o número total de irmãos e o divide pelo

número de pessoas. No caso, adicionamos  $0 + 1*1 + 1*2 + 5*2 = 13$ . Você tem cinco pessoas na turma, então a média é de  $13/5 = 2,6$ .

A *mediana* é o número no meio se você alinhá-los, do menor ao maior. Fariamos a linha 0, 1, 2, 5, 5. O terceiro número está no meio e, em nosso caso, significa que a mediana é dois.

Vemos que os três tipos de médias têm números diferentes. Quando usar um em vez do outro? A média é a mais comum, mas a mediana é útil se tiver pontos fora da curva. Suponha que uma pessoa tinha 1.000 irmãos! De repente, sua média fica muito maior, mas não representa realmente o número de irmãos que a maioria das pessoas tem. Por outro lado, a mediana permanece a mesma.

## Observações

É improvável que alguém sendo entrevistado para uma vaga de ciência de dados não saiba os diferentes tipos de médias. Então, essa questão está na verdade testando suas habilidades de comunicação em vez de saber se você acerta as definições (embora se você se enganar, é um sinal vermelho). Em nosso exemplo, usamos um bem simples que uma criança de oito anos poderia encontrar na vida real. Recomendamos manter o número de assuntos simples; não perca tempo fazendo cálculos para a média ou mediana porque está tentando calculá-las para 50 pontos de dados. Se houver um quadro branco na sala, talvez seja útil escrever os números para acompanhá-los. Como um bônus, você pode adicionar como fizemos, quando quiser usar um tipo de média em vez de outro.

### A.3.2 Explique o valor p

*Você pode me explicar o que é um valor p e como ele é usado?*

#### Resposta de exemplo

Imagine que você estava jogando cara ou coroa e virou 26 caras de 50 vezes. Você concluiria que a moeda tem alguma trapaça porque não virou 25 caras? Não! Você entende que a aleatoriedade está em jogo. Mas e se a moeda virou caras em 33 vezes? Como decidir qual é o limite para concluir que não é uma moeda sem trapaças?

É aqui que entra o *valor p*. Um *valor p* é a probabilidade de que, se a hipótese nula for verdadeira, veremos um resultado como ou mais extremo do que aquele que temos. Uma hipótese nula é nossa suposição-padrão, como nenhuma diferença entre dois grupos, que estamos tentando provar o contrário. Em nosso caso, a hipótese nula é que a moeda é sem trapanças.

Como um *valor p* é uma probabilidade, ele está sempre entre 0 e 1. O *valor p* é essencialmente uma representação de quão chocados ficaríamos com um resultado se nossa hipótese nula fosse verdadeira. Podemos usar um teste estatístico para calcular a probabilidade de que, se viramos uma moeda sem trapanças, obteríamos 33 ou mais caras ou coroas (ambas sendo resultados tão extremos quanto aquele que temos). Acontece que a probabilidade, o *valor p*, é 0,034. Por convenção, as pessoas usam 0,05 como o limite para rejeitar a hipótese nula. Nesse caso, rejeitaríamos a hipótese de que a moeda é sem trapanças.

Com um limite do *valor p* de 0,05, estamos aceitando que 5% do tempo, quando a hipótese nula é verdadeira, ainda vamos rejeitá-la. Essa é a nossa taxa falso-positiva: a taxa de rejeição da hipótese nula quando é realmente verdade.

## Observações

Essa questão está testando se você compreende o que um *valor p* é e pode comunicar a definição de maneira eficaz. Há equívocos comuns sobre o *valor p*, como o de que é a probabilidade de um resultado ser um falso-positivo. Ao contrário da questão de médias na seção anterior, é possível que alguém se engane. No lado da comunicação, recomendamos a utilização de um exemplo para orientar a explicação. Os cientistas de dados precisam ser capazes de se comunicar com uma grande variedade de stakeholders, alguns dos quais nunca ouviram falar de valores *p* e outros que pensam que entendem o que é, mas, na verdade, não sabem. Mostre que sabe o que são valores *p* e que consegue explicar aos outros.

### A.3.3 Explicar uma matriz de confusão

*O que é uma matriz de confusão? Para que pode utilizá-la?*



## Resposta de exemplo

Uma *matriz de confusão* permite que você veja como suas previsões se comparam com os resultados reais de determinado modelo. É uma grade 2x2 que tem quatro partes: o número de verdadeiro-positivos, falso-positivos, verdadeiro-negativos e falso-negativos. A partir de uma matriz de confusão, é possível calcular métricas diferentes, como precisão (a porcentagem classificada corretamente como verdadeiro-positivo ou verdadeiro-negativo) e sensibilidade, também conhecida como a taxa verdadeira-positiva, bem como a porcentagem de positivos classificados corretamente como tal. As matrizes de confusão são usadas em problemas de aprendizagem supervisionados, nos quais está classificando ou prevendo um resultado, como se um voo será atrasado ou se uma imagem é de um gato ou de um cachorro. Vamos tratar do exemplo dos resultados de voos.

	Atraso real	Horário real
Atraso previsto	60	15
Horário previsto	30	120

Nesse caso, 60 voos cuja previsão era atrasar, realmente se atrasaram, mas 30 previstos para ocorrerem no horário se atrasaram. Isso significa que nossa verdadeira taxa positiva é de  $60 / (60 + 30) 2/3$ .

Ver a matriz de confusão em vez de uma única métrica pode ajudá-lo a compreender melhor o desempenho do seu modelo. Digamos que, para um problema diferente, você acabou de calcular a precisão, por exemplo, e descobriu que tem 97% de precisão. Soa ótimo, mas poderia acabar que 97% dos voos ocorreram no horário. Se o modelo predisse simplesmente que cada voo está no horário, teria a exatidão de 97%, já que todos aqueles “no horário” são classificados corretamente, mas o modelo seria totalmente inútil!

## Observações

Essa questão testa se o candidato está familiarizado com os modelos de aprendizado supervisionado. Também testa se ele sabe diferentes maneiras de avaliar o desempenho dos modelos. Em nossa resposta, compartilhamos duas métricas que você poderia calcular a partir de uma matriz de confusão, mostrando que você entende como ela poderia ser usada, bem como um

caso em que ver toda a matriz em vez de apenas uma métrica seja útil.

### A.3.4 Interpretando modelos de regressão

*Como você interpretaria esses dois resultados de modelo de regressão, com base nos dados de entrada e do modelo? Esse modelo encontra-se em um conjunto de dados de 150 observações de três espécies de flores: setosa, versicolor e virginica. Para cada flor, são registrados o comprimento da sépala, a largura da sépala, o comprimento da pétala e a largura desta é uma regressão linear que prediz o comprimento da sépala das outras quatro variáveis.*

#### Dados de entrada no modelo

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa

#### Chamada de modelo

```
Model <- lm (Sepal.Length ~., iris)
```

#### Saída 1

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	2.17	0.280	7.76	1.43e-12
Sepal.Width	0.496	0.0861	5.76	4.87e- 8
Petal.Length	0.829	0.0685	12.1	1.07e-23
Petal.Width	-0.315	0.151	-2.08	3.89e- 2
Speciesversicolor	-0.724	0.240	-3.01	3.06e- 3
Speciesvirginica	-1.02	0.334	-3.07	2.58e- 3

#### Saída 2

variable	value
<chr>	<dbl>
r.squared	0.867
adj.r.squared	0.863
sigma	0.307
statistic	188

p.value	2.67e-61
df	6
logLik	-32.6
AIC	79.1
BIC	100
deviance	13.6
df.residual	144

## Resposta de exemplo

Olhando para os resultados apresentados no resumo do modelo, parece um modelo muito bom; o R-quadrado é 0,867, o que significa que os preditores explicam 86,7% da variância no comprimento da sépala. Os preditores são todos significativos no nível de  $p$  menor que é 0,05. Vejo que quanto mais larga for a sépala e mais longa for a pétala, mais longa será a sépala, enquanto as pétalas mais largas estão, na verdade, associadas a sépalas mais curtas. Tanto as espécies versicolor quanto a virginica têm coeficientes negativos, o que significa que seria possível prever que essas espécies tenham um comprimento menor do que a espécie setosa.

Suponha que encontramos uma nova flor, com uma sépala com largura 1, comprimento de 2 pétalas, pétala com largura 1 pétala, da espécie virginica. Nosso modelo preveria o comprimento da sépala da seguinte forma:  $2,17 + 0,496 * 1 + 0,829 * 2 - 0,315 * 1 - 1,02$ , que é de cerca de 3. Porém, antes de usar esse modelo, gostaria de ver mais alguns diagnósticos, como se os resíduos são normalmente distribuídos, e também de encontrar um conjunto de testes para ver como ele é executado de uma amostra para garantir que não esteja sobredimensionado.

## Observações

O entrevistador está procurando várias coisas, e você pode ganhar pontos, dependendo de quantos acertar. Nesse caso, o entrevistador quer confirmar se você entende estatística de modelo (como R-quadrado), bem como estimativas e valores  $p$  associados. Embora em nossa resposta essa informação não tenha sido explicitamente solicitada, acrescentamos como usaríamos esse modelo para prever o comprimento da sépala de uma nova flor. Por fim, adicionamos algumas informações sobre o modelo que gostaríamos de saber antes de começarmos a usá-lo. Esse tipo de pergunta

aberta é uma boa oportunidade para chegar ao ponto do que o entrevistador está provavelmente procurando e para adicionar a informação de bônus. Evite focar demais e acabar gastando 20 minutos em uma única questão; mostre que você compreende vários conceitos e siga adiante.

### **A.3.5 O que é impulsionalamento?**

*O que significa o termo “impulsionalamento” (boosting) ao referir-se a algoritmos de machine learning?*

#### **Resposta de exemplo**

*Impulsionalamento* refere-se a uma classe inteira de algoritmos de machine learning que são construídos com base em pegar um modelo fraco e reutilizá-lo vezes suficientes para que fique forte. A ideia é treinar um modelo fraco de dados, procurar áreas onde o modelo tenha erros e treinar um segundo modelo do mesmo tipo que pondere os pontos de dados onde ocorreram mais erros, esperando que o segundo modelo corrigirá alguns dos erros do primeiro. Repita esse processo repetidamente até atingir algum limite para o número de modelos. Depois, utilize todos esses modelos juntos para fazer a previsão. Com um grande conjunto de modelos, um resultado mais preciso será obtido do que se usasse um único modelo.

Uma implementação muito popular de um método de impulsionalamento é o XGBoost, o qual é bastante usado em linguagem R e Python.

#### **Observações**

*Impulsionalamento* é um termo incomum o suficiente para que seja de todo possível que alguém com um histórico básico de ciência de dados, possa não saber exatamente o que significa. Assim, essa questão é mais um teste para ver o nível de experiência do que de ter conhecimento em ciência de dados. A questão também é um pouco acadêmica; você pode imaginar alguém usando XGBoost com sucesso no código por anos sem pensar muito profundamente sobre como funciona. Essa questão é mais “é bom acertar, mas não é o fim do mundo caso não acerte” do que “se não acertar essa questão, é improvável que consiga a vaga”.

### **A.3.6 Algoritmo favorito**

*Qual é seu algoritmo de machine learning favorito? Você consegue explicá-lo para mim?*

– Jeroen Janssens

### **Resposta de exemplo**

Meu algoritmo favorito de machine learning é uma rede neural recorrente. Nos últimos tempos, tenho feito muitos trabalhos com processamento de linguagem natural, sendo que redes neurais recorrentes são ótimos modelos para classificar texto com rapidez.

Você conhece uma regressão linear? Uma rede neural é como uma regressão linear, exceto que há grupos de regressões lineares, sendo a saída de um grupo de regressões lineares a entrada para o próximo grupo. Ao unir todas essas regressões lineares em camadas de modelos, é possível fazer previsões com muito mais precisão.

Uma rede neural recorrente é um caso especial de uma rede neural sintonizada para dados que se enquadra em sequências. No caso do processamento de linguagem natural de um bloco de texto, a parte de saída por meio de uma sequência de palavras é a entrada para o modelo das palavras seguintes.

### **Observações**

Essa questão é uma das muitas que você pode receber durante uma entrevista que seja projetada para ver se pode explicar de maneira simples uma ideia complexa. O algoritmo escolhido para sua resposta não é tão importante quanto poder expressar como ele funciona com clareza. Dito isso, essa questão é uma excelente oportunidade para destacar os trabalhos anteriores interessantes que você tenha feito ao expressar um algoritmo que se relaciona com o trabalho e ao falar sobre ele.

### **A.3.7 Dados de teste x de treinamento**

*O que são dados de treinamento e o que são dados de teste? Qual é sua estratégia geral para criar esses conjuntos de dados?*

## Resposta de exemplo

Dados de treinamento são aqueles utilizados para treinar um modelo de machine learning. Os dados de teste são aqueles que não são utilizados no treinamento de um modelo de machine learning; em vez disso, são usados para validar o funcionamento do modelo. Esses conjuntos de dados têm de ser separados porque, se os dados forem utilizados para treinar um modelo, o modelo pode aprender o resultado correto para os dados e será artificialmente bom para ajustar-se a ele.

Existem muitas formas de separar os dados de treinamento dos dados de teste. Minha abordagem geral é pegar uma pequena amostra aleatória, como 10%, no início de uma análise e usá-la como dados de teste para todos os meus modelos, enquanto os outros 90% são dados de treinamento. Quando tiver encontrado um modelo que eu goste e que tenha um bom desempenho, retreino o modelo em todos os dados (treinamento e teste) para obter o modelo mais preciso para ser implantado na produção.

## Observações

É realmente importante ter uma boa explicação da diferença entre os dados de treinamento e os de teste, porque entender a distinção e como pensar sobre ela é uma parte fundamental da criação de um modelo de machine learning. Assim, existem muitas estratégias válidas para separar os dados. Além de amostragem aleatória, por exemplo, é possível utilizar a validação cruzada para evitar o enviesamento do seu modelo ao treinar em mais dados. Então, desde que tenha uma explicação lógica de por que você escolheu um método, deve dar tudo certo.

### A.3.8 Seleção de funcionalidades

*Como faria a seleção de funcionalidades se tivesse 1.000 covariáveis e precisasse de reduzi-las para 20?*

– Alex Hayes

## Resposta de exemplo

Existem várias formas diferentes de selecionar. Uma solução possível no caso de um problema de previsão é usar uma regressão de Lasso. Uma

regressão de Lasso é um tipo especial de regressão linear que aplica uma penalidade para aumentar o valor dos coeficientes. Ao aumentar o termo de penalidade na regressão, você pode fazer com que o modelo tenha cada vez menos coeficientes até usar apenas as 20 covariáveis mais importantes. Assim, o modelo seleciona quais coeficientes devem estar no modelo. Embora uma regressão de Lasso tenha um escore de precisão menor do que uma regressão linear com todas as covariáveis nos dados de treinamento, tem o benefício de usar apenas um pequeno número deles e pode ter melhor desempenho nos dados de teste, já que a Lasso reduz a probabilidade de sobredimensionamento.

Também é possível utilizar técnicas de redução dimensional como a Análise de Componentes Principais (PCA, do inglês) para reduzir a dimensão do problema de 1.000 para 20. A abordagem de Lasso escolherá 20 funcionalidades fora das 1.000 existentes. Métodos como a análise de componentes principais criarão 20 novas funcionalidades que tentam capturar o tanto dos dados dos 1.000 quanto lhe for possível.

## **Observações**

Existem muitas soluções possíveis para essa questão. Uma solução mais ampla é tentar usar uma função de etapa para remover covariáveis repetidamente até chegar a 20. Você poderia até mesmo pegar várias amostras de conjuntos de 20 funcionalidades e escolher aquele que funciona melhor. Pense nessa questão menos como um teste de saber a abordagem certa a um problema e mais como um teste de conseguir mostrar-lhe que conseguiria encontrar uma solução caso enfrentasse o problema. Essa questão é um teste para garantir que você não ficará travado no trabalho. Você pode pensar em alguma coisa que gostaria de tentar? Se sim, ótimo! Você pode tentar. Em caso negativo, você pode enfrentar dificuldades ao trabalhar sozinho.

Se forem dadas várias respostas, esteja preparado para responder à questão de seguimento: quando você usaria uma em vez da outra? Essa questão é uma maneira de verificar se você compreende as técnicas ou apenas as escolheu porque alguém lhe disse para usá-las. Nesse caso, você poderia responder que há uma vantagem entre a interpretação e a captura da

variabilidade: a de Lasso é facilmente interpretável, mas a PCA captura a maior variabilidade possível. A escolhida depende do que está buscando alcançar com a análise.

### **A.3.9 Implantação de um novo modelo**

*Você desenvolveu um novo modelo que tem um melhor desempenho do que o antigo modelo atualmente em produção. Como você determina se deve mudar o modelo na produção? Como abordar isso?*

– Emily Spahn

### **Resposta de exemplo**

Para mim, a resposta depende de alguns fatores no ambiente. Primeiro, por qual métrica o novo modelo funciona melhor? Partindo do princípio de que é uma precisão geral, verifico se o modelo é suficientemente melhor a ponto de valer a pena trocar o modelo antigo. Se for apenas um ponto percentual melhor na precisão, pode não valer a pena o esforço de mudar, porque o efeito pode ser insignificante. Em seguida, existe o risco de perturbar o modelo atual? Se o modelo foi implantado usando um pipeline bem mantido com log e testes claros, provavelmente faria a troca, mas se o modelo foi implantado movendo manualmente um modelo para um sistema de produção por alguém que não está mais na empresa, provavelmente eu me deteria.

Finalmente, existe uma forma de fazer um teste A/B no modelo primeiro? Preferencialmente, gostaria que os modelos antigos e novos fossem executados em paralelo para poder testar quaisquer problemas com o novo modelo ou casos de borda perdidos por ele. Nenhum sistema de teste pode abranger tudo desde a produção, por isso, seria ideal poder tê-lo em funcionamento para um conjunto selecionado de clientes ou entradas.

### **Observações**

A implantação de um modelo é, muitas vezes, uma proposta de risco e que dá bastante trabalho para uma empresa. Essa questão determina se você entende como isso é e como abordaria a situação. Um cientista de dados menos experiente ou engenheiro de machine learning pode sentir que a



escolha certa é implementar o modelo mais preciso o mais rapidamente possível, mas há riscos que precisam ser gerenciados. Se tiver alguma experiência com a qual possa se basear (como falhas nas implantações de modelos), essa questão é um ótimo momento para mencioná-la. Se não tiver, tudo bem; tente descrever o que pensa que pode sair mal.

### **A.3.10 Comportamento do modelo**

*Com base em um modelo que desenvolveu, como você projetaria uma métrica para avaliá-lo partindo da perspectiva do usuário final? Como você decidiria quais erros são aceitáveis?*

– Tereza Iofciu e Bertil Hatt

### **Resposta de exemplo**

Métricas de modelo-padrão, como o R-quadrado ou precisão, podem perder a perspectiva do usuário final ou do negócio. Um modelo de classificação poderia estar certo 99% do tempo, mas em 1% das vezes está errado, é um problema para o negócio a ponto de que o modelo nunca seria usado.

Acho que a melhor maneira de avaliar um modelo é tentar executar um experimento com ele. Se estou criando um modelo para agrupar clientes em segmentos, por exemplo, apresentaria os clusters ao marketing e os faria tentar executar um teste de marketing personalizado para um conjunto de amostra de clientes de diferentes segmentos. Eu compararia o desempenho do marketing com e sem os clientes segmentados, e, se houver uma melhoria significativa, o modelo é um sucesso. É totalmente diferente do uso de métricas sobre o próprio modelo, como a eficiência com que ele executa a segmentação, pois esses tipos de medidas analisam apenas o modelo. Aqui, estou de fato analisando o desempenho sem comparar com nenhum modelo.

A desvantagem da execução de um experimento com o modelo, sendo muitas vezes difícil configurar o experimento. Por vezes, não é possível dividir os clientes naqueles que obtêm o modelo e naqueles que não o fazem. Outras vezes, o efeito do modelo é tão pequeno que não apareceria em quaisquer KPIs que sejam fáceis de medir. Mas apesar dessas dificuldades, se for possível fazer um experimento, quase sempre é a

melhor abordagem.

## **Observações**

Essa questão é complicada porque é muito geral, mas, para respondê-la, você precisa falar sobre especificidades. Sua resposta pode variar drasticamente para um modelo preditivo em comparação a um modelo não supervisionado ou, se estiver trabalhando com marketing, em comparação ao departamento de operações. Fale muito sobre a ideia de que as medidas estatísticas não são as mesmas medidas com que a empresa se preocupa; os cientistas de dados inexperientes podem ficar excessivamente concentrados em maximizar as medidas estatísticas e ignorar as do negócio. Mas como você acaba falando sobre essas ideias, é muito aberto. Assim como em muitas respostas, se puder trazer exemplos de suas experiências, você pode agregar bastante profundidade.

### **A.3.11 Design do experimento**

(Questão, resposta e observações de Ryan Williams)

*Você está desenvolvendo um aplicativo e quer determinar se um layout recém-projetado seria melhor do que o atual. Como estruturaria um teste para escolher o melhor layout do aplicativo?*

### **Resposta de exemplo**

Há inúmeras formas diferentes de responder às especificidades dessa pergunta, mas os testes A/B geralmente seguem este tipo de fluxo:

1. Defina o que *melhor* significa escolhendo a(s) métrica(s) com as quais você se preocupa em melhorar: usuários ativos, cliques de botão, impressões e assim por diante.
2. Escolha uma hipótese nula baseada na sua métrica de sucesso, como: “Os cliques de botão serão os mesmos para todos os grupos”. Use essa hipótese para executar um cálculo de poder, o qual lhe dirá por quanto tempo você precisa executar o teste para detectar uma mudança de determinado tamanho.
3. Separe aleatoriamente sua população de usuários do aplicativo em grupos e forneça a cada grupo uma versão diferente do aplicativo.

4. Depois de executar o teste pelo período de tempo decidido na etapa 2, avalie se você vê uma diferença estatisticamente significativa entre os dois grupos usando um teste estatístico apropriado (como um teste t).

## **Observações**

Questões como essa são comuns para cargos de ciência de dados em equipes que estão muito envolvidas na medição da mídia, no desenvolvimento de apps/web e assim por diante. O entrevistador normalmente só quer saber se você compreende o propósito e os princípios gerais de testes A/B, especialmente para cargos mais iniciais. Em vez de focar nas especificidades dos testes de estatísticas (como quando usar um teste qui-quadrado em vez de um teste t), recomendamos definir uma abordagem clara de alto nível ao responder essa questão para demonstrar que você sabe como projetar uma experiência e determinar a causalidade.

### **A.3.12 Falhas no design do experimento**

(Questão, resposta e observações de Ryan Williams)

*Suponha que você tenha feito um teste A/B para selecionar um layout melhor do aplicativo; qual é um caso em que você pode não querer implementar o novo layout apesar de ver uma melhoria estatisticamente significativa na métrica que está testando?*

## **Resposta de exemplo**

Você não implementaria o layout se vir que ele afeta negativamente outras métricas importantes (métricas de proteção ou sem danos). Um exemplo pode ser uma situação em que a métrica em teste é de cliques por usuário e, embora você veja uma melhoria significativa em cliques por usuários expostos ao novo layout, também vê páginas no aplicativo levando mais tempo para carregar nesse layout. No caso, a degradação no desempenho do aplicativo pode não valer a pena o aumento em cliques, porque ao longo do tempo, uma pior experiência no aplicativo pode levar os usuários a se afastarem.

## **Observações**

Essa pergunta é muito aberta. O entrevistador quer ver se você reconhece que encontrar apenas um valor  $p$  baixo nem sempre é uma razão boa o suficiente para considerar uma experiência bem-sucedida. É arriscado uma empresa fazer alterações em um produto ao vivo, como um aplicativo ou site, e um único teste estatístico geralmente não encapsula todas as informações necessárias para tomar a decisão certa. Algumas outras respostas razoáveis a esse tipo de pergunta incluem a visualização de uma melhoria pequena demais em relação ao custo e o risco de alteração do aplicativo ou viés na metodologia de amostragem/separação.

### **A.3.13 Viés nos dados amostrados**

(Questão, resposta e observações de Ryan Williams)

*Que tipos de vieses você deve levar em conta ao usar dados de amostra? Como você pode saber se uma amostra tem viés?*

#### **Resposta de exemplo**

Muitos tipos de viés podem afetar os dados amostrados. Um dos vieses mais comuns em aplicações práticas de ciência de dados é o viés de seleção (selecionando sua amostra incorretamente). O viés de seleção pode ocorrer em cenários como na seleção de um grupo aleatório de clientes de uma tabela de nível de transação, que super-representa clientes com várias transações. Outros tipos de viés comuns incluem o de sobrevivência (a amostra super-representa um grupo que passou por algum processo de pré-seleção) e viés de resposta voluntária (a amostra super-representa um grupo que era mais provável de fornecer informação sobre si mesmo).

Existem métodos estatísticos que podem ser usados para identificar o viés em uma amostra, como comparar o valor médio de sua amostra com uma média conhecida ou esperada da população. Também pense racionalmente sobre o processo de amostragem para identificar vieses, tentando responder a esta questão: há algo sobre a maneira como temos amostrado este grupo que pode torná-lo diferente da população com que nos importamos?

#### **Observações**

Essa questão destina-se a testar sua compreensão sobre as limitações ao

trabalhar com dados e tirar conclusões. É menos importante entender termos específicos, como *viés de seleção* e *viés de sobrevivência*, do que entender as maneiras pelas quais os dados podem ser limitados ou mal-informados. O entrevistador quer ver se você compreende as nuances de trabalhar com dados do mundo real – todos enviesados de uma forma ou de outra – e toda confusão que esses dados implicam. O uso de dados de uma pesquisa opcional, por exemplo, tem um claro viés de resposta voluntária. Isso não significa que os dados sejam inutilizáveis, mas significa que você deve estar ciente do viés, pensar nas consequências que ele tem em sua análise e levá-lo em conta em quaisquer conclusões que fizer.

## **A.4 Comportamental**

### **A.4.1 Projeto que teve maior impacto**

*Qual é o projeto em que você trabalhou que teve o maior impacto?*

#### **Resposta de exemplo**

Em meu último trabalho, fui contratado para construir um sistema de análise de experimento online, ou um teste A/B. A empresa estava interessada em começar a executar experimentos e contava com um engenheiro que poderia implementá-los, dois encarregados do marketing de crescimento que traziam as ideias e estabeleciam as mudanças, além de um gerente, mas todos precisavam de uma maneira para entender quais eram os resultados do experimento.

Quando comecei, analisei cada experimento individualmente em linguagem R, mas sabia que este não era o melhor sistema: significava que a equipe precisava de mim para executar os scripts para ver os resultados e que eu estava duplicando o trabalho nas análises.

Foi isso que me levou a construir um dashboard interno para monitorar os experimentos. Esse dashboard incluiu não apenas os resultados de cada experimento, como a porcentagem de pessoas que se registraram ou se inscreveram no controle em comparação ao grupo de tratamento, mas também verificações de saúde para garantir que o experimento estivesse funcionando como esperado e que os resultados pudessem ser confiáveis.

Com esse dashboard, qualquer pessoa da empresa poderia ver os resultados mais atualizados.

Quando saí, esse dashboard estava sendo utilizado para todos os experimentos em andamento em cinco equipes. Graças ao trabalho que fiz com o resto da equipe de experimentos, quase todas as funcionalidades lançadas pela empresa são testadas pela primeira vez como um experimento para medir se têm algum impacto positivo.

## **Observações**

Para essa resposta, caso tiver feito qualquer projeto de ciência de dados para uma empresa, use um desses em vez de um projeto que não seja de ciência de dados. Por outro lado, se tiver feito um projeto de ciência de dados apenas para uso pessoal ou por conta de tarefas acadêmicas, você pode destacar outro projeto. O ponto principal aqui é focar no impacto no negócio. Dizer “construí um modelo com 90% de precisão!” não é o que estão buscando; eles querem entender como alguém usou o modelo, a ferramenta ou a análise que você construiu e por que foi importante.

### **A.4.2 Os dados surpreendem**

*Você pode contar sobre uma situação em que encontrou algo nos dados que o surpreendeu?*

#### **Resposta de exemplo**

Meu trabalho anterior era em uma empresa que ganhou dinheiro com assinaturas. Lá trabalhei em experimentos e, quando comecei, calculava a taxa de assinatura em um experimento como a porcentagem das pessoas que entraram e que fizeram a assinatura posteriormente. Embora pareça bom, revelou-se que as pessoas tinham assinaturas começando no futuro!

Depois de falar com o cientista de dados que possuía os dados de assinatura, descobri que essas assinaturas com datas de início futuras eram assinaturas que alguém tinha pausado. Por exemplo, pegue uma usuária com uma assinatura mensal começando em setembro. Ela poderia escolher pausar em vez de renovar ou cancelar a assinatura em outubro, não pagando e perdendo acesso por dois meses, mas recomeçar a assinatura em

dezembro. Nesse caso, ela teria duas linhas na tabela de assinaturas: uma para a assinatura de setembro a outubro e, depois, a partir de dezembro.

Para o meu caso de uso, não queria contar as assinaturas que começariam porque elas deixariam o status de pausa; queria apenas as assinaturas que alguém estivesse escolhendo ativamente.

Aprendi duas lições: que eu nunca deveria fazer suposições sobre os dados e que talvez precise personalizar uma fonte de dados para minhas necessidades. Assumi que seria impossível as assinaturas começarem no futuro, então, não havia verificado isso. Quando percebi esse problema, não sobrescrevi os dados originais, pois outras pessoas ainda precisavam saber sobre assinaturas que tinham sido definidas para começar no futuro. Em vez disso, fiz minha própria tabela que contava apenas com novas assinaturas.

## **Observações**

Nessa resposta, usamos um exemplo no qual ficamos surpreendidos com o que é essencialmente uma questão de qualidade de dados para nosso caso de utilização. Mas você poderia falar sobre uma situação na qual sua intuição simplesmente não coincidiu com os resultados, por exemplo, uma análise exploratória de dados que você fez de uma subthread do Reddit sobre ciência de dados, quando pensou que a contagem de palavras dos posts se correlacionaria de modo positivo com o número de comentários, mas acabou sendo uma correlação negativa. Você também quer se certificar de que explica por que teve sua suposição inicial.

Essa questão está testando se você pensa sobre seus dados antes de simplesmente mergulhar neles. Também está testando que você não tenta apenas confirmar sua hipótese inicial, mas se deixa surpreender pelos resultados e a se adaptar à nova informação.

### **A.4.3 Reflexões sobre trabalhos anteriores**

*Qual é a coisa que você mais queria mudar em seu trabalho anterior que não conseguia?*

– Bertil Hatt

## **Resposta de exemplo**

Descobri que, em minha última empresa, havia dificuldades reais de comunicação. A equipe de liderança estava constantemente pedindo que as pessoas fossem mais abertas e expressassem suas preocupações, mas isso não acontecia. Minha teoria é que dizia respeito aos próprios líderes não estarem abertos; eles constantemente nos diziam que tudo estava indo muito bem quando sabíamos que havia problemas.

Algo que eu queria muito modificar era fazer a liderança abrir-se mais para nós. Se eles expressassem mais suas próprias dificuldades e preocupações, isso teria facilitado aos funcionários mais inexperientes abrirem-se e terem criado um ambiente de trabalho melhor.

## **Observações**

Essa questão é *delicada*! Você precisa mostrar que compreendeu seu ambiente de trabalho anterior suficientemente bem para ter uma proposta de melhoria para ele, mas precisa fazê-lo soando como se tivesse um bom relacionamento com seu empregador anterior.

Você pode listar vários tipos diferentes de alterações, como alterações técnicas, alterações na dinâmica de equipe e alterações nos produtos. Quanto mais significativa uma mudança que possa listar, melhor (desde que não seja: “Queria que tivesse mais refrigerantes gratuitos”). Também é ótimo se puder refletir sobre o motivo pelo qual essa mudança não aconteceu (“Queria que tivéssemos usado uma linguagem moderna como R ou Python, mas estávamos usando SAS por causa de todos os produtos herdados que mantínhamos”). Explicar por que a mudança não ocorreu mostra que você colocou o pensamento nas limitações do ambiente.

Evite insultar seu empregador anterior (“Acredita que eles não eram inteligentes o suficiente para usar FORTRAN?!”). Não passe a impressão de que um dia você deixará a próxima empresa e falará mal dela também. Seja respeitador do trabalho que seu antigo empregador fazia, mesmo que apresentasse falhas.

### **A.4.4 Pessoas experientes cometendo um erro baseado em dados**

*O que faria se tivesse cálculos ou resultados que conflitassem com os*



*resultados anteriores de uma pessoa experiente na empresa? Você tentaria convencê-los de que estava certo e, em caso afirmativo, como?*

– Hlynur Hallgrímsson e Heather Nolis

## **Resposta de exemplo**

Em primeiro lugar, eu me questionaria se esse resultado é suficientemente importante para ser levantado. Se ocorreu por uma pequena porcentagem, mas ainda tomaríamos a mesma decisão com os novos resultados ou se os resultados anteriores nunca fossem usados para nada, eu poderia deixar assim mesmo.

Se não, começaria tentando entender as motivações e os objetivos da outra pessoa. Suponha que essa pessoa era o vice-presidente de vendas e fizeram uma análise que mostra que cada vendedor contratado rendeu em mais de duas vezes o salário nas vendas. Por conta disso, eles usaram essa análise para justificar a contratação de mais cinco pessoas para a equipe. Se eu mostrar que cada vendedor, na verdade, rende menos que o salário, comprometeria todo o departamento de vendas. As pessoas teriam muito em jogo com esse resultado, por isso seria importante ter cuidado.

Eu agendaria uma reunião com essas pessoas. Ao compreender a situação, seria possível fazer uma suposição educada sobre como reagiriam. Se os resultados conflitavam por terem cometido um erro na análise ou se o resultado fosse fundamental para o negócio, eu esperaria que ficassem na defensiva e tentassem encontrar falhas na minha análise, então, me prepararia emocionalmente e conferiria várias vezes meus resultados. Tentaria, ainda, encontrar uma solução que permitisse que eles se salvassem, mudaria a estratégia e colocaria o negócio na direção certa.

No pior dos casos – se não escutassem nem apresentassem motivos válidos de por que os novos resultados estavam errados, e eu achasse que os novos resultados são vitais para o negócio – gostaria de trabalhar com meu gerente para apresentar uma estratégia para que os novos resultados sejam compartilhados e colocados em prática. Infelizmente, às vezes, as pessoas não concordarão com sua nova análise, e o foco precisa passar de convencê-las a encontrar uma forma de atingir seus objetivos e limitar o impacto de uma análise errada.

## Observações

Essa questão busca compreender como você lidaria com conflitos com alguém mais experiente. Embora existam muitas respostas, algumas delas como: “Enviaria um email a todos os membros da empresa para falarem publicamente sobre como estavam errados” ou “Trataria sempre deste modo, independentemente da situação, pois a única coisa importante são os dados” constituiria, sem dúvida, um problema. Os acadêmicos podem especialmente ter dificuldades em lidar com conflitos nas empresas; no meio acadêmico, palestras podem ser um concurso de quem, no público, pode encontrar mais falhas na pesquisa e derrubar os argumentos. Em empresas, por outro lado, você precisa ser capaz de compartilhar seu ponto de vista e ajudar o negócio a tomar decisões certas, ao mesmo tempo em que coloca na balança outros fatores e compreende as nuances de diferentes situações. O entrevistador está à procura de sinais de que você tenha resolvido com sucesso discordâncias anteriores, razão pela qual deve apresentar essa experiência tanto quanto possível.

### A.4.5 Discordâncias com colegas da equipe

*Fale-me de uma situação na qual discordou de um colega da equipe. Do que se tratava e o que você fez?*

### Resposta de exemplo

Uma vez, estava trabalhando com um gerente de produto em um experimento no qual seria definido um tempo de execução de duas semanas com base em um cálculo de poder. Quatro dias depois, ele queria suspender o experimento antes e lançá-lo em sua totalidade porque o valor  $p$  era 0,04 na principal métrica de sucesso. Mas eu sabia que poderia ser algo para dar uma rápida olhada: verificar os resultados diariamente para ver se o valor  $p$  cai abaixo de 0,05 e parar quando acontece, aumenta-se extremamente a taxa de falso-positivo. Também sabia que o gerente de produto estava muito incentivado a ter um experimento bem-sucedido: uma das principais métricas em que foram avaliados foi a receita incremental obtida com experimentos bem-sucedidos.

Nesse caso, foquei em certificar-me de que sabia de onde estavam vindo e

em fazer perguntas. Lembrei dos nossos objetivos compartilhados: tornar a empresa o mais bem-sucedida possível. Expliquei um exemplo simples do webcomic xkcd para ajudá-los a desenvolver a intuição sobre por que parar cedo poderia ser um problema: que se você verificar se 20 cores diferentes de jujubas seriam associadas à acne, mesmo que nenhuma fosse, por um acaso um teste estatístico provavelmente “encontraria” uma associação. (Aqui está o link para os quadrinhos, para referência: <https://xkcd.com/882/>) Da mesma forma, estávamos perseguindo fantasmas estatísticos e ficamos suscetíveis de nos enganarmos a pensar que tínhamos um impacto positivo quando não era o caso. No fim, concordaram em manter o experimento funcionando durante as duas semanas planejadas.

Essa situação também levou-me a pensar mais sobre como eu poderia melhorar a ferramenta de experimento para facilitar que as pessoas fizessem a coisa certa. Uma plataforma de experimento que conheço tem um círculo pequeno que se enche mais a cada dia e torna-se uma marca de seleção ao final de sete dias. Isso ajudou as pessoas a executarem seus experimentos durante, pelo menos, uma semana, que é a melhor prática.

## **Observações**

Essa resposta utiliza a abordagem STAR (situação, tarefa, abordagem, resultado), a qual é uma estrutura clássica para responder questões de entrevista comportamental, pois fornece uma estrutura para a resposta que é fácil de seguir. Quando se pensa em um bom exemplo para essa questão, encontre uma situação que teve um resultado positivo, e não “E depois nunca mais nos falamos” ou “fui demitido”. O desacordo deve estar relacionado com o trabalho em si, não algo como “não concordamos com a forma de encher a máquina de lavar louça do escritório”. Os entrevistadores estão buscando ver se você consegue ter empatia com alguém com quem discorda e evitar falar mal deles ou culpá-los pelo seu problema.

### **A.4.6 Problemas difíceis**

*O que você faz quando não sabe como resolver um problema relacionado com a ciência de dados?*

## **Resposta de exemplo**

Para perguntas de codificação, o Google é meu amigo! Muitas vezes, uma resposta à pergunta do Stack Overflow é o primeiro resultado se procurar no Google uma mensagem de erro ou algo do tipo: “Como faço a alocação latente de Dirichlet em linguagem R?”. Se eu souber qual é a função ou o pacote que quero usar, mas não tenho certeza de como ele funciona, vou verificar se há alguma documentação disponível.

Mas, às vezes, não sei como abordar um problema. Nesses casos, normalmente começo a quebrar o problema, às vezes escrevendo os diferentes componentes no quadro branco. Isso me ajuda a focar nas questões centrais, que podem ser aquelas que sei como resolver, mesmo que inicialmente o problema todo parecesse assustador.

Também gosto da regra de gastar 15 ou 30 minutos no problema (dependendo se sinto que estou tendo algum progresso) e, em seguida, pedir ajuda a outro cientista de dados da empresa. É minha responsabilidade tentar descobrir isso sozinho primeiro, mas também o fato de não ficar preso em algo durante um dia inteiro, quando um colega poderia ter me ajudado em alguns minutos. Quando eu entro em contato, compartilho o que tentei junto com um exemplo pequeno e reproduzível para facilitar que a outra pessoa veja o problema (em vez de enviar centenas de linhas de código para análise).

## **Observações**

A ciência de dados é um campo no qual você continuamente estará aprendendo e sendo desafiado por problemas que nunca viu antes; portanto, é importante desenvolver algumas estratégias quando ficar travado. Uma coisa que essa questão está buscando é que você tenha desenvolvido estratégias fora do ambiente da sala de aula, onde havia uma folha de respostas, colegas e um professor para ajudá-lo. Você poderá precisar adaptar essa resposta à empresa com a qual está falando. Se disser que sua estratégia principal é perguntar aos seus colegas cientistas de dados, e estiver sendo entrevistado para ser o primeiro cientista de dados, essa resposta será um sinal vermelho.

## **A.5 Problemas de lógica**

### A.5.1 Estimativa

*Qual seria uma estimativa de quantos minifrascos de xampu são utilizados por todos os hotéis dos Estados Unidos em um ano?*

#### Resposta de exemplo

Faço o cálculo do número de frascos utilizando a seguinte fórmula:

número de hotéis nos EUA \* número médio de quartos por hotel \* 1 frasco de xampu por quarto ocupado por noite \* utilização média do quarto \* 365 dias por ano = número de frascos de xampu por ano

Depois estimo os números na fórmula:

- *Número de hotéis nos Estados Unidos* – se assumir que existe um hotel para cada 5.000 pessoas no país, e há cerca de 300 milhões de pessoas no país, ou seja, 60.000 hotéis.
- *Número de quartos por hotel* – cinquenta parece uma suposição razoável para o número médio de quartos em um hotel daqueles em que já me hospedei.
- *Utilização média do quarto* – como os hotéis precisam ser lucrativos, acredito um quarto tem uma chance de 80% de estar ocupado por noite.

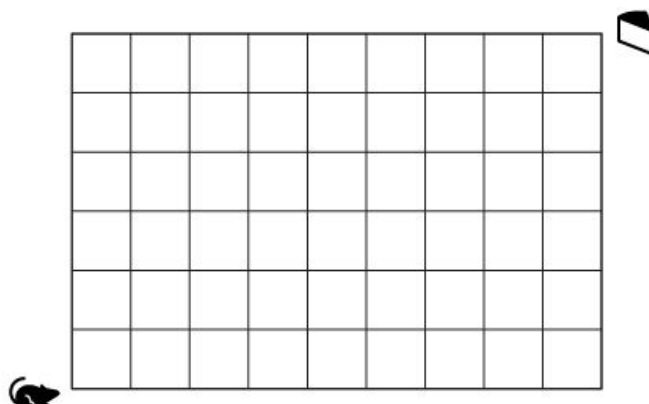
A fórmula fica  $60.000 * 50 * 1 * 0,8 * 365 = 876$  milhões de frascos.

#### Observações

A solução para essa questão é criar uma fórmula para o número que você está tentando estimar e supor os números a serem colocados na fórmula. Existem muitas versões dessa questão, de “Quantas bolas de tênis de mesa podem caber num Boeing 747?” a “Quantos pianos há na França?”. O entrevistador está buscando verificar se você pode pensar em uma fórmula que faça algum sentido e que sua lógica para adivinhar cada um dos números na fórmula também faça sentido. Quase não há chances de que você obtenha um número próximo ao certo durante a entrevista (não temos ideia se 50 é uma boa suposição para o número médio de quartos de hotel, por exemplo), mas isso não é importante.

Não há muito a fazer para se preparar para essas questões, exceto praticar o componente improvisacional de improvisar fórmulas e estimativas.

## A.5.2 Combinatórias



*Imagine uma grade como a ilustrada acima, com um rato no canto inferior esquerdo da grade. No canto superior direito, há um pedaço de queijo. O rato só pode andar nas linhas da grade e nunca se afastar dela. Quantos caminhos existem do rato ao queijo?*

### Resposta de exemplo

Para chegar ao queijo, o rato deve andar um espaço apenas na linha horizontal da grade nove vezes e, depois, andar um espaço ao longo de uma linha vertical na grade seis vezes (porque a grade tem 9x6). Vamos chamar o movimento horizontal de H e o movimento vertical de V. Então, qualquer string com 9 Hs e Vs é um caminho válido do início ao fim. Ir direto para cima e depois para a direita, por exemplo, seria VVVVVVHHHHHHHHH. Existem 15 maneiras fatoriais ( $15!$ ) de organizar 15 caracteres distintos, que são chamados de *permutações*, mas como 6 deles são a mesma letra (V) e 9 deles são a mesma letra (H), temos de remover todos os arranjos duplicados. Podemos removê-los contando quantos duplicados de cada existem. Os Vs são duplicados  $6!$  vezes (o número de maneiras que podem ser arranjados) e os Hs são duplicados  $9!$  vezes. Isso significa que a resposta é  $15!/(6!)(9!)$ , ou 5.005 caminhos.

### Observações

Essa questão é realmente difícil de responder. Primeiro, é difícil saber a resposta certa. Se tiver estudado de alguma forma o campo das combinatórias, você pode conhecê-lo; caso contrário, é difícil de repente

dar-se conta de que é possível tanto pensar no problema como arranjar caminhos. Mesmo que veja essa maneira de formular o problema, você pode não saber contar o número de soluções.

Em segundo lugar, mesmo que conheça a resposta, é difícil explicar o problema de uma forma clara e dar a solução sem se estender demais. Não suponha que todos conhecem termos como *permutação*, mas, se fosse explicar tudo, você passaria muito tempo nisso.

Por fim, não há uma maneira de estudar para essa questão. Há tantas questões de combinatórias que você não pode ter respostas preparadas antecipadamente para todas. Sua melhor aposta para questões como essas é explicar seu processo de pensamento e como você pode abordar o problema. Se o entrevistador está dando muito peso a questões como essa, é um sinal vermelho!

O'REILLY®



# Introdução ao GraphQL

BUSCA DE DADOS COM ABORDAGEM DECLARATIVA PARA APLICAÇÕES WEB MODERNAS

novatec

Eve Porcello & Alex Banks



# Introdução ao GraphQL

Porcello, Eve

9788575227107

216 páginas

[Compre agora e leia](#)

Por que o GraphQL é a tecnologia mais inovadora para buscar dados desde o Ajax? Ao oferecer uma linguagem de consulta para suas APIs e um runtime para responder às consultas com seus dados, o GraphQL representa uma alternativa ao REST e às arquiteturas ad hoc dos web services. Com este guia prático, Alex Banks e Eve Porcello apresentam um caminho de aprendizado objetivo aos desenvolvedores web de frontend, engenheiros de backend e gerentes de projeto e de produto que queiram começar a trabalhar com o GraphQL. Você explorará a teoria dos grafos, a estrutura de dados de grafo e os tipos do GraphQL antes de aprender a construir um esquema para uma aplicação de compartilhamento de fotos na prática. Este livro também apresenta o Apollo Client: um framework popular que pode ser usado para conectar o GraphQL à sua interface de usuário. - Explore a teoria dos grafos e analise exemplos conhecidos de grafos em uso nos dias de hoje. - Saiba como o GraphQL aplica métodos de consulta de banco de dados à internet. - Crie um esquema para uma

aplicação PhotoShare, que servirá como um roteiro e um contrato entre as equipes de frontend e de backend. - Use JavaScript para implementar um serviço GraphQL totalmente funcional, e o Apollo para implementar um cliente. - Aprenda a preparar APIs GraphQL e os clientes para um ambiente de produção

[Compre agora e leia](#)

# CANDLESTICK

Um método para ampliar lucros na Bolsa de Valores



novatec

Carlos Alberto Debastiani

# Candlestick

Debastiani, Carlos Alberto

9788575225943

200 páginas

[Compre agora e leia](#)

A análise dos gráficos de Candlestick é uma técnica amplamente utilizada pelos operadores de bolsas de valores no mundo inteiro. De origem japonesa, este refinado método avalia o comportamento do mercado, sendo muito eficaz na previsão de mudanças em tendências, o que permite desvendar fatores psicológicos por trás dos gráficos, incrementando a lucratividade dos investimentos.

Candlestick – Um método para ampliar lucros na Bolsa de Valores é uma obra bem estruturada e totalmente ilustrada. A preocupação do autor em utilizar uma linguagem clara e acessível a torna leve e de fácil assimilação, mesmo para leigos. Cada padrão de análise abordado possui um modelo com sua figura clássica, facilitando a identificação. Depois das características, das peculiaridades e dos fatores psicológicos do padrão, é apresentado o gráfico de um caso real aplicado a uma ação negociada na Bovespa. Este livro possui, ainda, um índice resumido dos padrões para pesquisa rápida na utilização cotidiana.

[Compre agora e leia](#)

Marcos Abe

# MANUAL DE ANÁLISE TÉCNICA

ESSÊNCIA E ESTRATÉGIAS AVANÇADAS

TUDO O QUE UM INVESTIDOR PRECISA SABER PARA  
PROSPERAR NA BOLSA DE VALORES ATÉ EM TEMPOS DE CRISE

novatec

# Manual de Análise Técnica

Abe, Marcos

9788575227022

256 páginas

[Compre agora e leia](#)

Este livro aborda o tema Investimento em Ações de maneira inédita e tem o objetivo de ensinar os investidores a lucrarem nas mais diversas condições do mercado, inclusive em tempos de crise. Ensinará ao leitor que, para ganhar dinheiro, não importa se o mercado está em alta ou em baixa, mas sim saber como operar em cada situação. Com o Manual de Análise Técnica o leitor aprenderá: - os conceitos clássicos da Análise Técnica de forma diferenciada, de maneira que assimile não só os princípios, mas que desenvolva o raciocínio necessário para utilizar os gráficos como meio de interpretar os movimentos da massa de investidores do mercado; - identificar oportunidades para lucrar na bolsa de valores, a longo e curto prazo, até mesmo em mercados baixistas; um sistema de investimentos completo com estratégias para abrir, conduzir e fechar operações, de forma que seja possível maximizar lucros e minimizar prejuízos; - estruturar e proteger operações por meio do gerenciamento de capital. Destina-se a iniciantes na bolsa

de valores e investidores que ainda não desenvolveram uma metodologia própria para operar lucrativamente.

[Compre agora e leia](#)





AVALIANDO  
EMPRESAS

# INVESTINDO EM AÇÕES

A APLICAÇÃO PRÁTICA DA  
ANÁLISE FUNDAMENTALISTA NA  
AVALIAÇÃO DE EMPRESAS

novatec

CARLOS ALBERTO DEBASTIANI  
FELIPE AUGUSTO RUSSO

# Avaliando Empresas, Investindo em Ações

Debastiani, Carlos Alberto

9788575225974

224 páginas

[Compre agora e leia](#)

Avaliando Empresas, Investindo em Ações é um livro destinado a investidores que desejam conhecer, em detalhes, os métodos de análise que integram a linha de trabalho da escola fundamentalista, trazendo ao leitor, em linguagem clara e acessível, o conhecimento profundo dos elementos necessários a uma análise criteriosa da saúde financeira das empresas, envolvendo indicadores de balanço e de mercado, análise de liquidez e dos riscos pertinentes a fatores setoriais e conjunturas econômicas nacional e internacional. Por meio de exemplos práticos e ilustrações, os autores exercitam os conceitos teóricos abordados, desde os fundamentos básicos da economia até a formulação de estratégias para investimentos de longo prazo.

[Compre agora e leia](#)

O'REILLY



# Microserviços prontos para a produção

CONSTRUINDO SISTEMAS PADRONIZADOS EM UMA  
ORGANIZAÇÃO DE ENGENHARIA DE SOFTWARE



novatec

Susan J. Fowler

# Microserviços prontos para a produção

Fowler, Susan J.

9788575227473

224 páginas

[Compre agora e leia](#)

Um dos maiores desafios para as empresas que adotaram a arquitetura de microserviços é a falta de padronização de arquitetura – operacional e organizacional. Depois de dividir uma aplicação monolítica ou construir um ecossistema de microserviços a partir do zero, muitos engenheiros se perguntam o que vem a seguir. Neste livro prático, a autora Susan Fowler apresenta com profundidade um conjunto de padrões de microserviço, aproveitando sua experiência de padronização de mais de mil microserviços do Uber. Você aprenderá a projetar microserviços que são estáveis, confiáveis, escaláveis, tolerantes a falhas, de alto desempenho, monitorados, documentados e preparados para qualquer catástrofe. Explore os padrões de disponibilidade de produção, incluindo: Estabilidade e confiabilidade – desenvolva, implante, introduza e descontinue microserviços; proteja-se contra falhas de dependência. Escalabilidade e desempenho – conheça os componentes essenciais para alcançar mais eficiência do microserviço. Tolerância a falhas

e prontidão para catástrofes – garanta a disponibilidade forçando ativamente os microsserviços a falhar em tempo real.

Monitoramento – aprenda como monitorar, gravar logs e exibir as principais métricas; estabeleça procedimentos de alerta e de prontidão. Documentação e compreensão – atenuar os efeitos negativos das contrapartidas que acompanham a adoção dos microsserviços, incluindo a dispersão organizacional e a defasagem técnica.

[Compre agora e leia](#)