

# Projetos de Ciência de Dados com Python

**Abordagem de estudo de caso para a criação de projetos  
de ciência de dados bem-sucedidos usando Python,  
pandas e scikit-learn**

**Stephen Klosterman**

**Packt**  
Novatec

Copyright © Packt Publishing 2019. First published in the english language under the title ‘Data Science Projects with Python – (9781838551025)’

Copyright © Packt Publishing 2019. Publicação original em inglês intitulada ‘Data Science Projects with Python – (9781838551025)’

Esta tradução é publicada e vendida com a permissão da Packt Publishing.

© Novatec Editora Ltda. [2019].

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: Aldir Coelho Corrêa da Silva

Revisão gramatical: Tássia Carvalho

Editoração eletrônica: Carolina Kuwabata

ISBN do impresso: 978-65-86057-10-2

ISBN do ebook: 978-65-86057-11-9

Histórico de impressões:

Maio/2020                      Primeira edição

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110

02460-000 – São Paulo, SP – Brasil

Tel.: +55 11 2959-6529

E-mail: [novatec@novatec.com.br](mailto:novatec@novatec.com.br)

Site: [www.novatec.com.br](http://www.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)

GRA20200427

# Sumário

<b>Prefácio .....</b>	<b>7</b>
<b>Capítulo 1 ■ Exploração e limpeza de dados.....</b>	<b>11</b>
Introdução.....	11
Python e o sistema de gerenciamento de pacotes Anaconda .....	12
A indexação e o operador slice.....	13
Exercício 1: Examinando o Anaconda e familiarizando-se com o Python .....	15
Diferentes tipos de problemas da ciência de dados .....	18
Carregando os dados do estudo de caso com o Jupyter e o pandas.....	20
Exercício 2: Carregando os dados do estudo de caso em um Jupyter Notebook .....	22
Familiarizando-se com os dados e executando sua limpeza.....	25
O problema da empresa.....	26
Etapas da exploração de dados .....	27
Exercício 3: Verificando a integridade básica dos dados .....	28
Máscaras booleanas .....	32
Exercício 4: Continuando a verificação da integridade dos dados .....	34
Exercício 5: Explorando e limpando os dados .....	38
Exploração e garantia da qualidade dos dados.....	43
Exercício 6: Explorando o limite de crédito e as características demográficas.....	44
Aprofundamento nas características categóricas .....	48
Exercício 7: Implementando a OHE para uma característica categórica .....	51
Explorando as características de histórico financeiro do dataset .....	55
Atividade 1: Explorando as características financeiras restantes do dataset.....	63
Resumo .....	65
<b>Capítulo 2 ■ Introdução ao Scikit-Learn e avaliação do modelo.....</b>	<b>67</b>
Introdução.....	67
Examinando a variável de resposta e concluindo a exploração inicial .....	68
Introdução ao scikit-learn .....	71
Gerando dados sintéticos .....	77
Dados para uma regressão linear .....	77
Exercício 8: Regressão linear com o scikit-Learn.....	79
Métricas de desempenho de modelos para a classificação binária.....	82
Dividindo os dados: conjuntos de treinamento e de teste.....	83
Acurácia da classificação .....	86
Taxa de verdadeiros positivos, taxa de falsos positivos e matriz de confusão .....	88

Exercício 9: Calculando as taxas de verdadeiros e falsos positivos e negativos e a matriz de confusão em Python.....	90
Descobririndo probabilidades previstas: como a regressão logística faz previsões? .....	94
Exercício 10: Obtendo probabilidades previstas a partir de um modelo de regressão logística	94
Curva receiver operating characteristic (ROC).....	99
Precisão .....	103
Atividade 2: Executando a regressão logística com uma nova característica e criando uma curva precision-recall.....	104
Resumo .....	105

### **Capítulo 3 ■ Detalhes da regressão logística e exploração de características ..... 107**

Introdução.....	107
Examinando os relacionamentos entre as características e a resposta .....	108
Correlação de Pearson.....	111
Teste F .....	115
Exercício 11: Teste F e seleção de características univariada .....	116
Pontos mais importantes do teste F: equivalência com o teste t para duas classes e cuidados	120
Hipóteses e próximas etapas .....	121
Exercício 12: Visualizando o relacionamento entre as características e a resposta .....	122
Seleção de características univariada: o que ela pode ou não fazer.....	129
Entendendo a regressão logística com sintaxe de funções Python e a função sigmóide .....	130
Exercício 13: Plotando a função sigmóide .....	133
Escopo das funções.....	136
Por que a regressão logística é considerada um modelo linear? .....	138
Exercício 14: Examinando a conveniência das características para a regressão logística .....	141
Dos coeficientes da regressão logística às previsões com o uso da função sigmóide.....	144
Exercício 15: Limite de decisão linear da regressão logística .....	145
Atividade 3: Ajustando um modelo de regressão logística e usando os coeficientes diretamente..	153
Resumo .....	154

### **Capítulo 4 ■ O trade-off entre viés e variância..... 155**

Introdução.....	155
Estimando os coeficientes e as intercepções da regressão logística.....	156
Gradiente descendente para a descoberta de valores de parâmetros ótimos.....	158
Exercício 16: Usando o gradiente descendente para reduzir a função custo.....	161
Suposições da regressão logística .....	165
Motivação para a regularização: O trade-off entre viés e variância.....	170
Exercício 17: Gerando e modelando dados de classificação sintéticos .....	172
Regularização lasso (L1) e ridge (L2) .....	176
Validação cruzada: Selecionando o parâmetro de regularização e outros hiperparâmetros.....	181
Exercício 18: Reduzindo o overfitting no problema de classificação de dados sintéticos .....	187
Opções da regressão logística no scikit-learn .....	198
Escalonamento de dados, pipelines e características de interação no scikit-learn .....	199
Atividade 4: Validação cruzada e engenharia de características com os dados do estudo de caso.....	201
Resumo .....	204

**Capítulo 5 ■ Árvores de decisão e florestas aleatórias..... 205**

Objetivos do aprendizado .....	205
Introdução.....	205
Árvores de decisão .....	206
Terminologia das árvores de decisão e conexões com o Machine Learning .....	207
Exercício 19: Uma árvore de decisão no scikit-learn .....	209
Treinando árvores de decisão: Impureza dos nós.....	216
Características usadas para as primeiras divisões: Conexões com a seleção de características univariada e as interações .....	221
Treinando árvores de decisão: Um algoritmo ganancioso .....	221
Trenando árvores de decisão: Diferentes critérios de parada.....	222
Usando árvores de decisão: Vantagens e probabilidades previstas.....	223
Abordagem mais conveniente da validação cruzada .....	226
Exercício 20: Encontrando hiperparâmetros ótimos para uma árvore de decisão .....	228
Florestas aleatórias: Combinações de árvores de decisão.....	234
Floresta aleatória: Previsões e interpretabilidade .....	237
Exercício 21: Ajustando uma floresta aleatória .....	238
Gráfico quadriculado (checkerboard) .....	243
Atividade 5: Busca em grade na validação cruzada com floresta aleatória .....	245
Resumo .....	246

**Capítulo 6 ■ Imputação de dados faltantes, análise financeira e distribuição para o cliente..... 247**

Objetivos do aprendizado .....	247
Introdução.....	247
Revisão dos resultados dos modelos.....	248
Lidando com dados faltantes: Estratégias de imputação.....	250
Preparando amostras com dados faltantes .....	253
Exercício 22: Limpando o dataset.....	253
Exercício 23: Imputação de PAY_1 pela moda e aleatória .....	257
Um modelo preditivo para PAY_1.....	264
Exercício 24: Construindo um modelo de classificação multiclasse para a imputação .....	266
Usando o modelo de imputação e comparando-o com outros métodos.....	271
Confirmando o desempenho do modelo com o conjunto de teste desconhecido .....	275
Análise financeira .....	277
Conversa financeira com o cliente.....	278
Exercício 25: Caracterizando custos e economias.....	279
Atividade 6: Derivando insights financeiros.....	285
Considerações finais sobre a distribuição do modelo preditivo para o cliente.....	286
Resumo .....	288

**Apêndice..... 290**

Capítulo 1: Exploração e limpeza de dados.....	290
Atividade 1: Explorando as características financeiras restantes do dataset.....	290
Capítulo 2: Introdução ao Scikit-Learn e avaliação do modelo .....	296

Atividade 2: Executando a regressão logística com uma nova característica e criando uma curva precision-recall.....	296
Capítulo 3: Detalhes da regressão logística e exploração de características .....	300
Atividade 3: Ajustando um modelo de regressão logística e usando os coeficientes diretamente..	300
Capítulo 4: O trade-off entre viés e variância .....	303
Atividade 4: Validação cruzada e engenharia de características com os dados do estudo de caso	303
Capítulo 5: Árvores de decisão e florestas aleatórias.....	307
Atividade 5: Busca em grade na validação cruzada com floresta aleatória .....	307
Capítulo 6: Imputação de dados faltantes, análise financeira e distribuição para o cliente.....	311
Atividade 6: Derivando insights financeiros .....	311