

UNTITLED

Ana Luzielma
Jaylhane Nunes
Raianny Soares

07/02/2022

INTRODUÇÃO

CONTEXTUALIZAÇÃO

Motivadas pelo interesse comum em leitura optamos por realizar a análise de um conjunto de dados envolvendo livros.

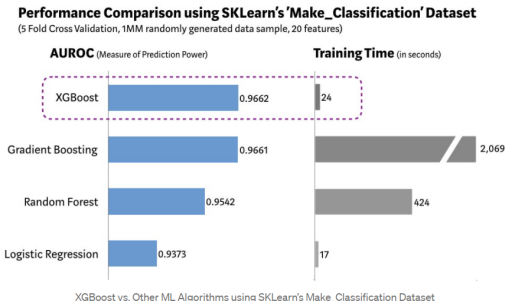
O conjunto de dados selecionado possui **11.131 observações**, foi originalmente obtido por meio de **raspagem** de dados na **API** da plataforma **GoodReads** e se encontra disponível no site **Kagle**.

Nele é possível encontrar as seguintes colunas:

bookID	average_rating	language_code	text_reviews_count
title	isbn	num_pages	publication_date
authors	isbn13	ratings_count	publisher

Chegamos em um consenso que geralmente há fatores influenciando na satisfação com a leitura e quisemos investigar se, com os dados disponíveis, seria possível obter um modelo que conseguisse prever se o livro foi considerado: *ruim*, *bom* ou *ótimo*.

UMA IDEIA INICIAL



Dentre as leituras realizadas sobre as possibilidades de modelos e métodos, percebemos que o XGboost tinha um ótimo desempenho comparado a outros modelos, e que, apesar da variável `average_rating` ser uma variável contínua um método de classificação poderia ser adequado para os nossos objetivos, desde que gerássemos categorias e encaixássemos os intervalos.

A INSPIRAÇÃO FINAL

Julia Silge

[ABOUT](#) [BLOG](#)

Tune xgboost models with early stopping to predict shelter animal status

By Julia Silge in [rstats](#) [tidymodels](#)

AUGUST 7, 2021

This is the latest in my series of [screencasts](#) demonstrating how to use the [tidymodels](#) packages, from just getting started to tuning more complex models. I participated in this week's episode of the [SLICED](#) playoffs, a competitive data science streaming show, where we competed to predict the status of shelter animals. 🐾 I used xgboost's early stopping feature as I competed, so let's walk through how and when to try that out!



INDICAMOS



juliasilge.com

► *Agora, ao nosso modelo!*

Julia Silge

DATA SCIENTIST & SOFTWARE ENGINEER



I'm a tool builder, author, international keynote speaker, and real-world practitioner focusing on data analysis and machine learning. I love making beautiful charts and communicating about technical topics with diverse audiences.

[READ LATEST POSTS →](#)

DESENVOLVIMENTO

ENGENHARIA DE DADOS

A análise exploratória consistiu em:

- ▶ Limpeza dos Dados
- ▶ Análise Descritiva

Dado os nossos objetivos, percebemos que algumas colunas eram dispensáveis e outras poderiam ser transformadas, de forma que:

Excluídas	Transformadas	Geradas
bookID	average_rating	book_rating ¹
title	language_code	
isbn	text_reviews_count	prop_text_reviews ²
authors	publication_date	
isbn13		
publisher		

Nosso MAIOR DESAFIO DURANTE A ANÁLISE E A MODELAGEM
ESTEVE RELACIONADO A ESSA FASE DE ENGENHARIA DE DADOS

ANÁLISE EXPLORATÓRIA

- ▶ Para iniciar separamos o conjunto em **treino** e **teste** baseando-nos em 75% das observações e balanceando-os com nossa variável resposta `book_rating`.

```
set.seed(1904, kind = "Mersenne-Twister", normal.kind = "Inversion")
livros_split <- initial_split(livros, prop = .75, strata = book_rating)
livros_treino <- training(livros_split)
livros_teste <- testing(livros_split)
```

- ▶ Em seguida verificamos a dispersão e correlação entre as variáveis numéricas:

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'GGally':
```

```
##   method from
```

```
##   +.gg      ggplot2
```

```
library(treino %>%
```