Price Prediction Model for Books

Presented by Kate Popkova

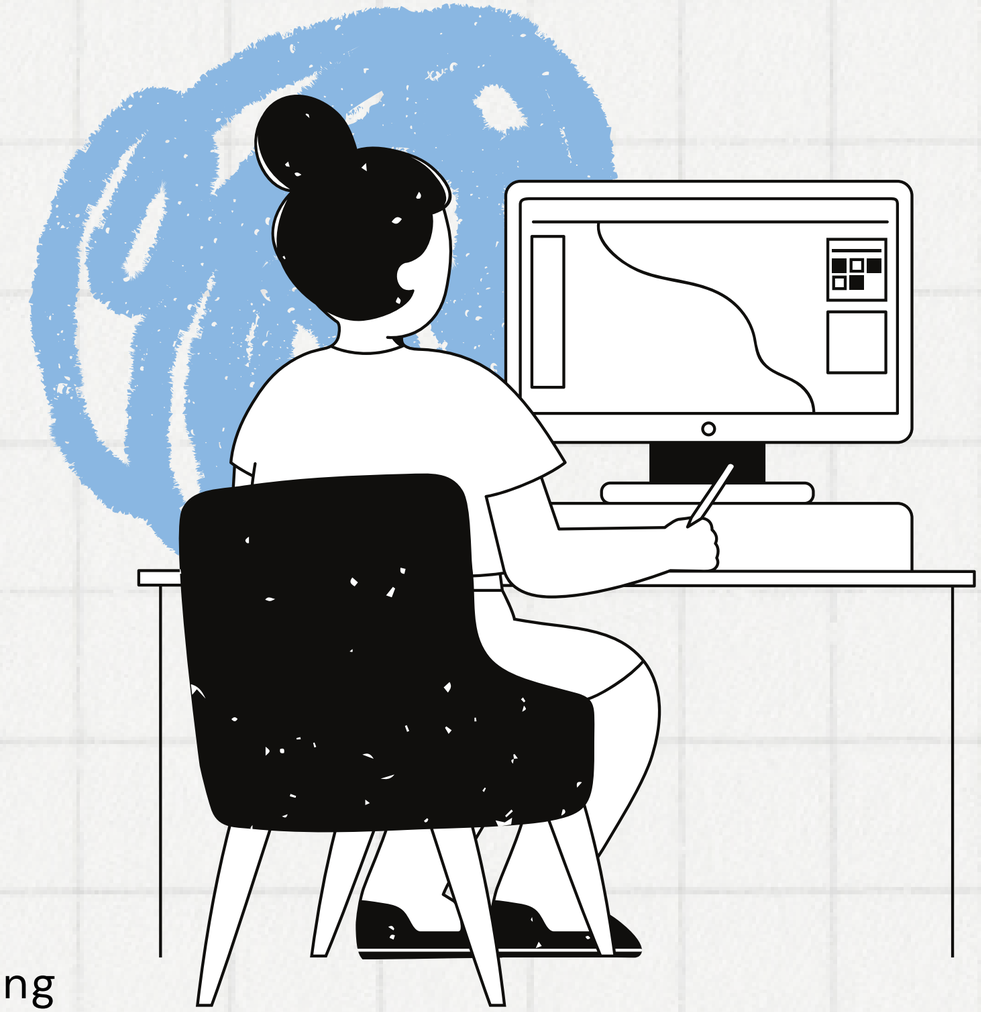
Problem Statement

The book industry faces the complex challenge of pricing books in a way that maximizes profit while remaining attractive to readers.

Factors influencing book prices include

- author popularity,
- genre trends,
- publication date, and
- market demand.

This creates an opportunity to create a Dynamic Pricing Strategy.



Who Experiences These Problems?

Publishers

Struggle to set optimal prices that balance profitability with market competitiveness.

Retailers

Both online and physical bookstores need to competitively price books to maximize sales and manage inventory.

Authors

Especially self-published authors, who must decide on the pricing of their books without the resources of large publishing houses.

The dataset I'll be using has 52,478 records on books collected from an e-commerce platform called **Iberlibro**.

Can Data Science help?

Machine learning can revolutionize book pricing by providing a predictive model that considers a wide range of factors influencing a book's value in the market.

Societal and Business Value

- More accurate pricing leads to **better sales** and profit margins.
- Reduces the likelihood of overstocking or understocking.
- Properly priced books become more **accessible** to a broader audience, potentially increasing literacy and education.

The dataset was created by two students in **2020** in Barcelona, Spain.

There are **25 columns** and **52,478 rows.**

There are some missing values, notably in '**edition**', '**series**', and '**firstPublishDate**' columns.

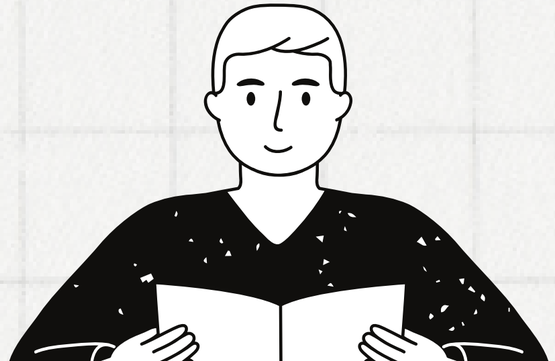
The dataset also has **50 duplicated rows**, but no duplicated columns.

There are **20 categorical** columns and **5 numerical** columns.

Book cover **images** are provided within the dataset.

Dataset

Goodreads Best Books Ever



Attributes	Definition
bookId	Book Identifier as in goodreads.com
title	Book title
series	Series Name
author	Book's Author
rating	Global goodreads rating
description	Book's description
language	Book's language
isbn	Book's ISBN
genres	Book's genres
characters	Main characters
bookFormat	Type of binding
edition	Type of edition (ex. Anniversary Edition)
pages	Number of pages
publisher	Editorial
publishDate	Publication date
firstPublishDate	Publication date of first edition
awards	List of awards
numRatings	Number of total ratings
likedPercent	Derived field, percent of ratings over 2 starts (as in GoodReads)
setting	Story setting
coverImg	URL to cover image
bbeScore	Score in Best Books Ever list
bbeVotes	Number of votes in Best Books Ever list
price	Book's price in EUR (extracted from Iberlibro)

Next Steps

01

Clean the dataset

- Remove the duplicated rows
- Handle the missing values
- Converting data types
- Identify outliers and decide what to do with them

02

Visualize the dataset

- Explore the distribution of numerical and categorical values.

03

Image Feature Extraction

- Research methods and find out if the image can correlate with the book's genre, target audience, or popularity, which might indirectly affect the price.

04

Select a Model

- Decide what model to use to predict book prices.

Thank you very much!

https://github.com/katepopkova8/my_capstone_project