

Price Prediction Model for Books

Sprint 3

Presented by Kate Popkova



Non-technical overview

Problem Statement: The book industry's challenge is optimally pricing books to balance profitability with market appeal. Key stakeholders are publishers, retailers, and authors.

Proposed Solution: A machine learning model for predicting book prices, considering factors like author popularity, genre trends, and market demand.

Potential Impact:

- Increased sales and profit margins, competitive pricing, efficient stock management.
- Societal Value: Greater book accessibility, promoting literacy, equitable compensation for authors.



★ Dataset and Preprocessing Procedures

Dataset Description

52,478 book records from IberLibro, including various features influencing book prices.

Preprocessing Steps

- **Data Cleaning:** Handling missing values, outliers.
- **Feature Engineering:** Creating new variables for better insights.
- **Data Normalization:** Scaling features for model compatibility.

New Features

Book Series (Binary)

Edition (Binary)

Genres (Top 20)

Publishers (Top 20)

Book Formats

Pages (Short/Medium/Long)

Target Variable

Price



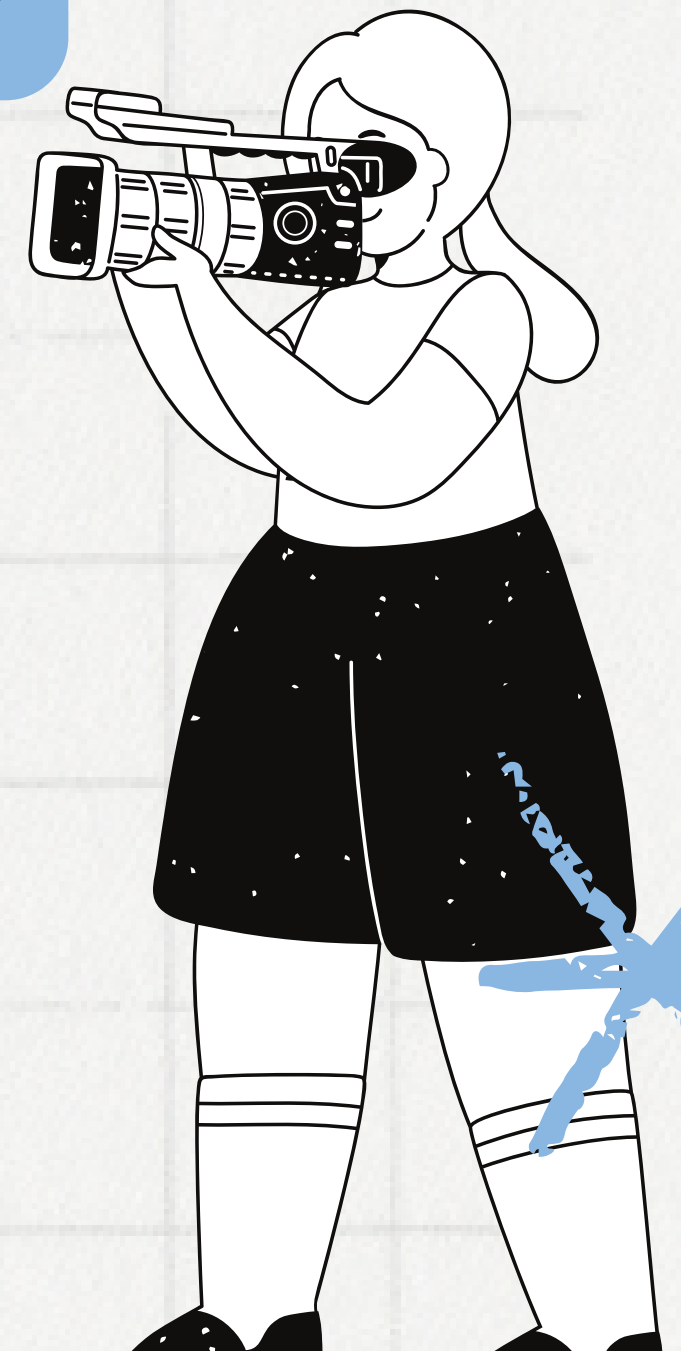
Baseline Models

- Linear Regression
- Decision Tree
- Random Forest



Advanced Models

- Gradient Boosting Machines (GBM)
- Ridge Regression
- Lasso Regression



Evaluation Metrics

- Mean Squared Error (MSE)
- R-squared value
- Model comparison



Results

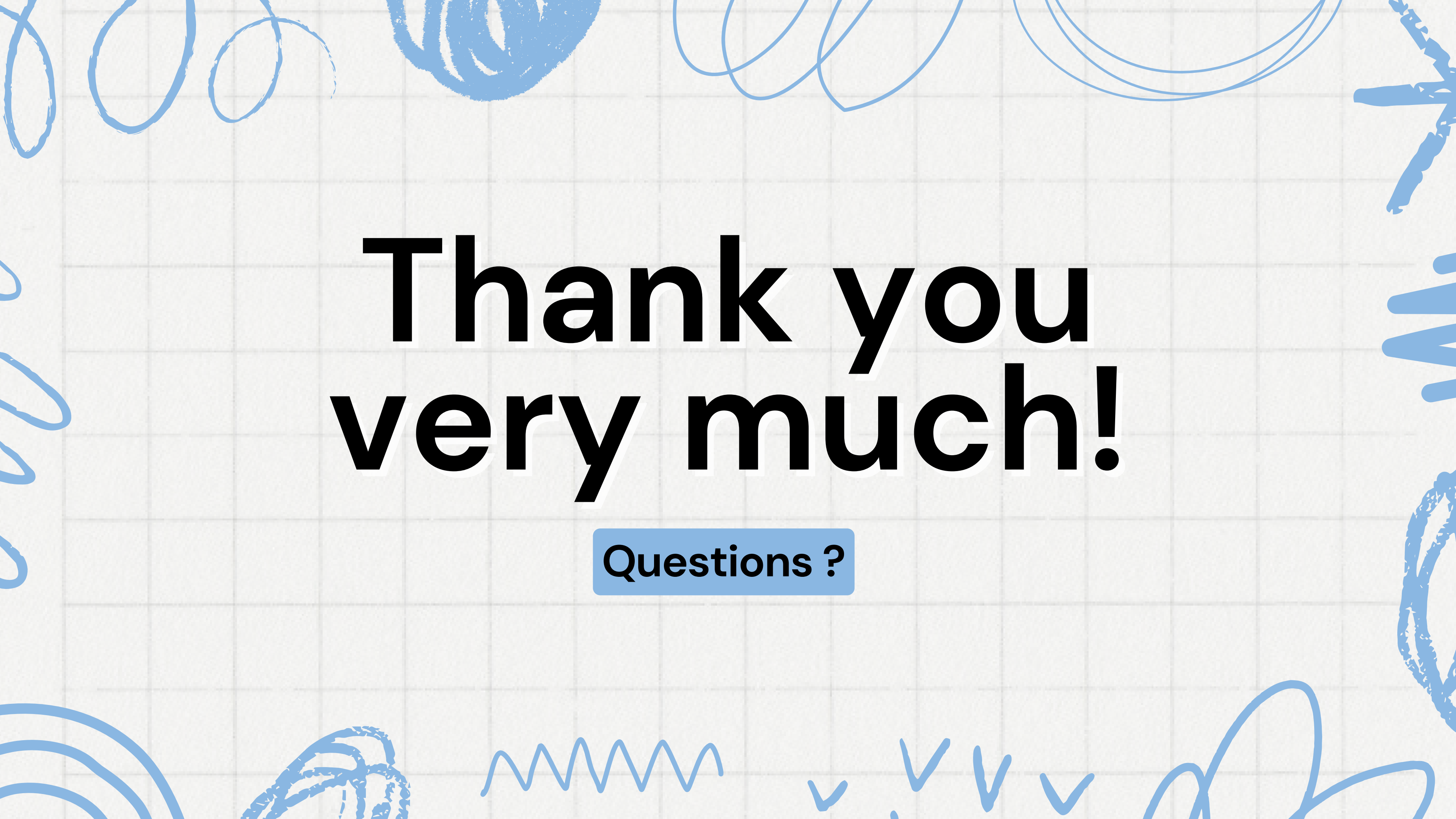
MODEL	MSE	MAE	R-SQUARED
Linear Regression	415.02	6.47	0.020
Decision Tree	426.89	6.52	-0.007
Decision Tree (Tuned)	417.11	6.46	0.015
Random Forest	426.37	6.46	0.016
GBM	409.18	6.32	0.033
Ridge Regression	414.93	6.47	0.020
Lasso Regression	421.50	6.60	0.005

Streamlit Demo

Local URL:

<http://localhost:8503/>



The background is a light gray grid. It is decorated with various hand-drawn blue doodles, including circles, loops, a star-like shape, a zigzag line, and several checkmarks.

Thank you very much!

https://github.com/katepopkova8/my_capstone_project