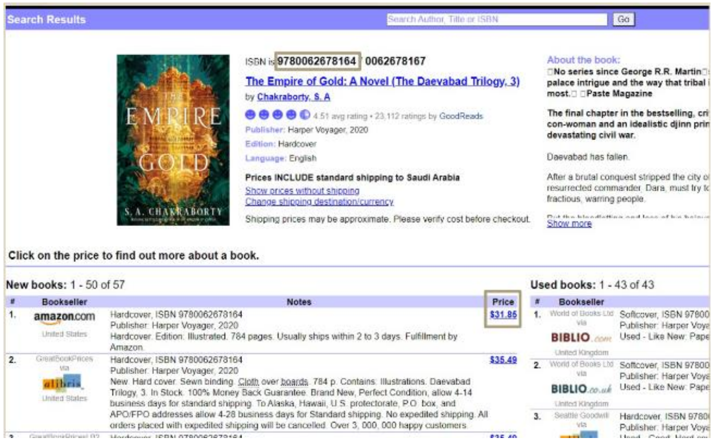
Introduction

This is the second project for the T5 Data Science Bootcamp, which is about building a linear regression model that address a useful prediction using Scikit-learn, and using data scraped from website with Requests, BeautifulSoup, and Selenium. This project helps ReadMe firm which is an online web site that is selling books. In order to determine the price of its products they ask for our help as data scientists to study the books features to predict a suitable price.

Design And Data

we aim to produce a model that is able to investigate book features such as title, authors, average_rating, isbn, isbn13, language_code, num_pages, ratings_count, text_reviews_count, publication_date, and publisher. In order to determine a suitable price.
We collect the data from two deferent data sources, first source is a CSV dataset publicly available at: https://www.kaggle.com/jealousleopard/goodreadsbooks it has originally more than 11,000 rows and deferent 12 columns. We select 1298 rows and 12 columns.
The second source is the price of these books we gain it from https://www.bookfinder.com we use web scrap selenium to fetch this information. Taking advantage of selenium feature that allow the browser to interact with the web site specially in our case we make selenium searching for book prices by its isbn. Then we save the results into csv file and merge it with our previous dataset.



ISBN is 9780062678164 0062678167

The Empire of Gold: A Novel (The Daevabad Trilogy, 3)

by Chakraborty, S. A

4.51 avg rating • 23,112 ratings by GoodReads

Publisher: Harper Voyager, 2020

Edition: Hardcover

Language: English

Prices INCLUDE standard shipping to Saudi Arabia

Show prices without shipping
Change shipping destination/currency

Shipping prices may be approximate. Please verify cost before checkout.

About the book:

No series since George R.R. Martin palace intrigue and the way that tribal most. Paste Magazine

The final chapter in the bestselling, cri con-woman and an idealistic djinn prin devastating civil war.

Daevabad has fallen.

After a brutal conquest stripped the city o resurrected commander, Dara, must try to fractious, warring people.

Show more

Click on the price to find out more about a book.

New books: 1 - 50 of 57

#	Bookseller	Notes	Price
1.	amazon.com United States	Hardcover, ISBN 9780062678164 Publisher: Harper Voyager, 2020 Hardcover. Edition: Illustrated. 784 pages. Usually ships within 2 to 3 days. Fulfillment by Amazon.	$31.86
2.	GreatBookPrices via alibris United States	Hardcover, ISBN 9780062678164 Publisher: Harper Voyager, 2020 New. Hard cover. Sewn binding. Cloth over boards. 784 p. Contains: Illustrations. Daevabad Trilogy, 3. In Stock. 100% Money Back Guarantee. Brand New, Perfect Condition, allow 4-14 business days for standard shipping. To Alaska, Hawaii, U.S. protectorate, P.O. box, and APO/FPO addresses allow 4-28 business days for Standard shipping. No expedited shipping. All orders placed with expedited shipping will be cancelled. Over 3, 000, 000 happy customers	$35.49

Used books: 1 - 43 of 43

#	Bookseller	
1.	World of Books Ltd via BIBLIO.com United Kingdom	Softcover, ISBN 97800 Publisher: Harper Voya Used - Like New: Pape
2.	World of Books Ltd via BIBLIO.co.uk United Kingdom	Softcover, ISBN 97800 Publisher: Harper Voya Used - Like New: Pape
3.	Seattle Goodwill via	Hardcover, ISBN 9780 Publisher: Harper Voya

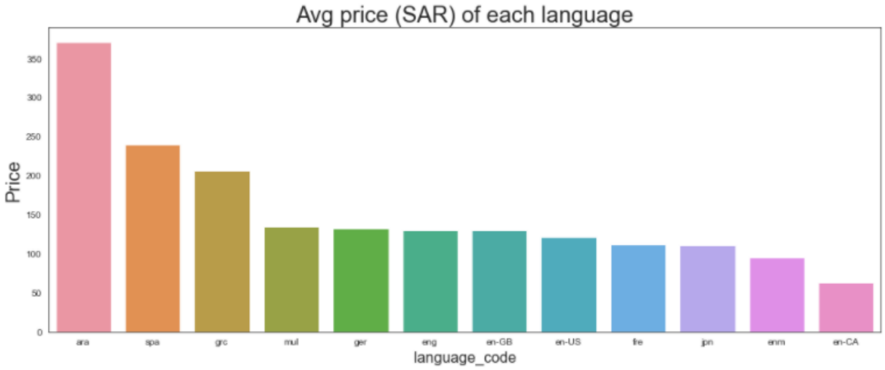
Algorithms

- We build a linear regression in python, and techniques such as regularization and polynomial features, adding interaction terms and dummy variables have been used.
- Rigorous model selection and evaluation has been used to select model between Linear Regression, Ridge Regression, Polynomial Regression, and Lasso Regression

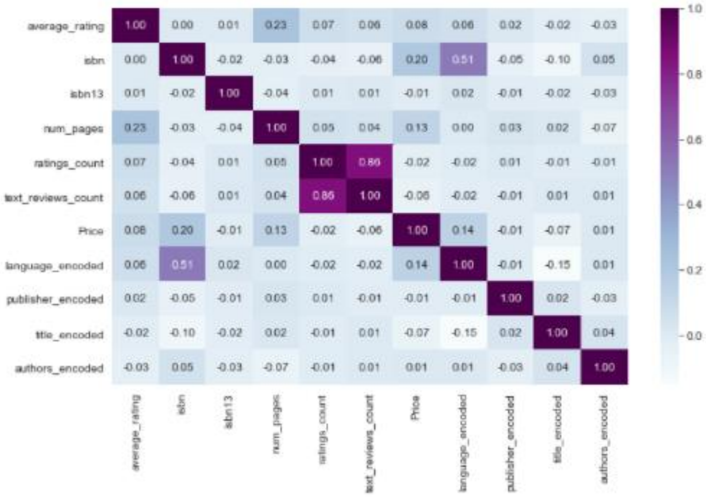
Tools

- selenium in python (Scraping the web)
- Pandas and Numpy (Exploring the data)
- Matplotlib and Seaborn (Visualizing the data)
- Sci-kit Learn (linear Regression model and other models)

Communication

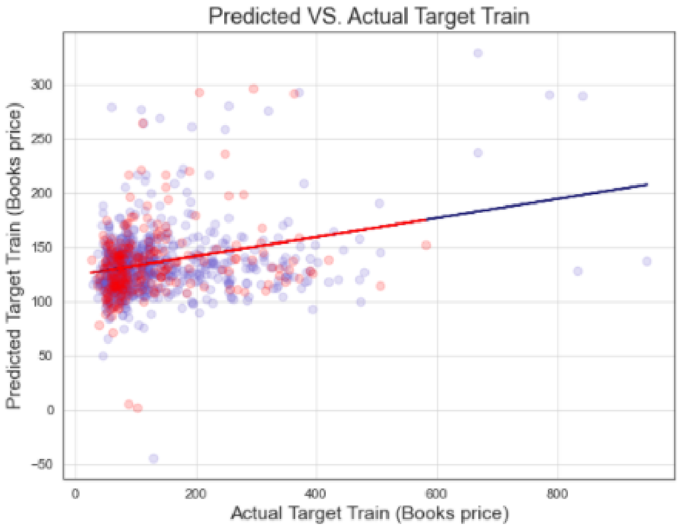


- Avg price (SAR) of each language

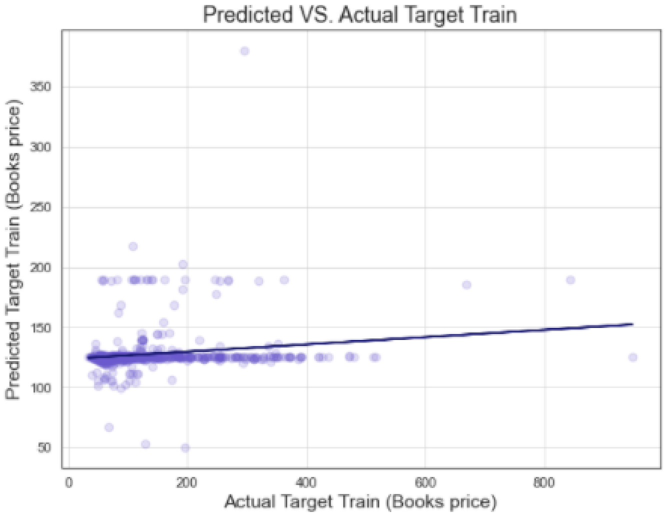


	average_rating	isbn	isbn13	num_pages	ratings_count	text_reviews_count	Price	language_encoded	publisher_encoded	title_encoded	authors_encoded
average_rating	1.00	0.00	0.01	0.23	0.07	0.06	0.08	0.06	0.02	-0.02	-0.03
isbn	0.00	1.00	-0.02	-0.03	-0.04	-0.06	0.20	0.51	-0.05	-0.10	0.05
isbn13	0.01	-0.02	1.00	-0.04	0.01	0.01	-0.01	0.02	-0.01	-0.02	-0.03
num_pages	0.23	-0.03	-0.04	1.00	0.05	0.04	0.13	0.00	0.03	0.02	-0.07
ratings_count	0.07	-0.04	0.01	0.05	1.00	0.86	-0.02	-0.02	0.01	-0.01	-0.01
text_reviews_count	0.06	-0.06	0.01	0.04	0.86	1.00	-0.06	-0.02	-0.01	0.01	0.01
Price	0.08	0.20	-0.01	0.13	-0.02	-0.06	1.00	0.14	-0.01	-0.07	0.01
language_encoded	0.06	0.51	0.02	0.00	-0.02	-0.02	0.14	1.00	-0.01	-0.15	0.01
publisher_encoded	0.02	-0.05	-0.01	0.03	0.01	-0.01	-0.01	-0.01	1.00	0.02	-0.03
title_encoded	-0.02	-0.10	-0.02	0.02	-0.01	0.01	-0.07	-0.15	0.02	1.00	0.04
authors_encoded	-0.03	0.05	-0.03	-0.07	-0.01	0.01	0.01	0.01	-0.03	0.04	1.00

correlation between features



Model before feature engineering



Model after feature engineering

Cross Validation

Regression Algorithm	Linear Regression	Ridge Regression	Polynomial Regression (degree = 2)	Lasso Regression
Training Score	0.075843	0.044664	0.068357	0.066383
Validation Score	0.029533	0.00066	0.006979	0.002216