



ReadMe

LINEAR REGRESSION PROJECT REPORT

Norah Alqahtani, Batoul Alosaimi,
Shoroq Almotairi

SDAIA Academy | Data Science Boot Camp



Introduction

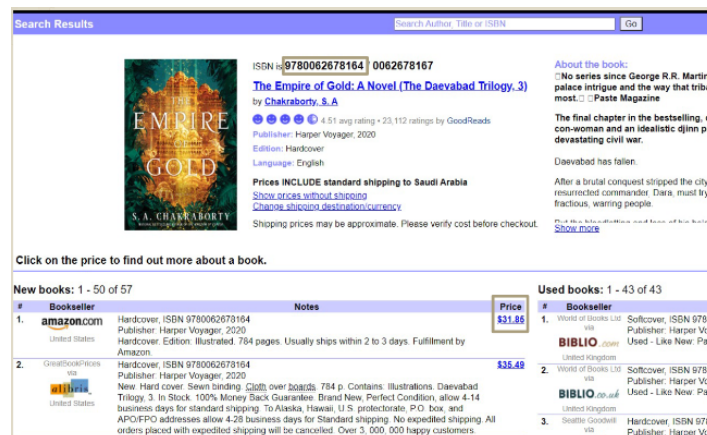
This is the second project for the T5 Data Science Bootcamp, which is about building a linear regression model that address a useful prediction using Scikit-learn, and using data scraped from website with Requests, BeautifulSoup, and Selenium. This project helps ReadMe firm which is an online web site 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 Description

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.



Search Results

ISBN: 9780062678164 0062678167

The Empire of Gold: A Novel (The Daevabad Trilogy, 3)
by Chakrabarty, 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's palace intrigue and the way that tribal most..." Paste Magazine
The final chapter in the bestselling, critically-acclaimed and an idealistic djinn print devastating civil war.
Daevabad has fallen.
After a brutal conquest stripped the city of resurrected commander, Dara, must try to frantically, warning 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	Hardcover, ISBN 9780062678164 Publisher: Harper Voyager; 2020 Hardcover, Edition: Illustrated, 784 pages. Usually ships within 2 to 3 days. Fulfillment by Amazon.	\$31.85
2.	GreatBookPrices	Hardcover, ISBN 9780062678164 Publisher: Harper Voyager; 2020 New, Hard cover, Seven 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	Notes	Price
1.	World of Books Ltd	Softcover, ISBN 9780062678164 Publisher: Harper Voyager Used - Like New, Paper	
2.	World of Books Ltd	Softcover, ISBN 9780062678164 Publisher: Harper Voyager Used - Like New, Paper	
3.	Seattle Goodwill	Hardcover, ISBN 9780062678164 Publisher: Harper Voyager	

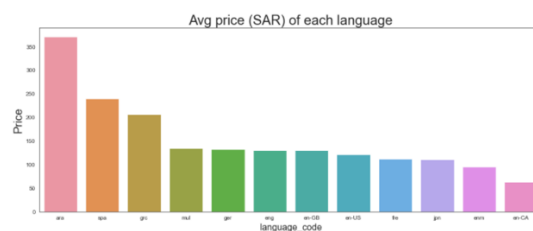
Algorithms

- 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

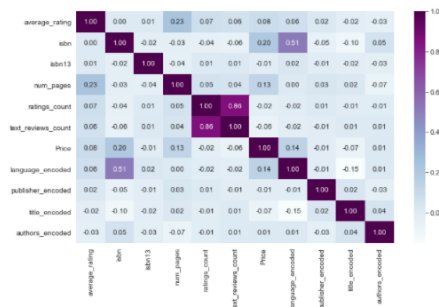
Tools

- 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)

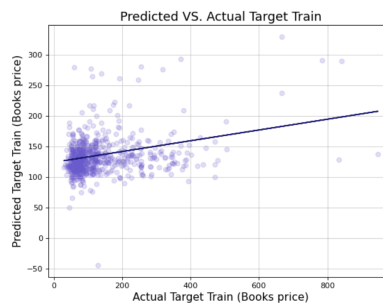
Communication



- Avg price (SAR) of each language



correlation between features



Compare actual Y with predicted Y in first learner regression

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

Calinges

- Choosing topic
 - We have been changed project topic 3 times due to finding data problem, even though there were a lot of datasets available online but in order to achieve project requirement that ask us to use scraping tools, so we need to scrap data instead of use datasets only. We find that website name 'book finder' contain books prices so we scrap it and merge it with our dataset.
- Scraping Data
 - Since we need to scrap data from several pages, and our original data set had more than 11,000 data point means our selenium will browse 11,000 pages; we struggled to gather that much of data because the web site blocked us many times, so we decided decrease data amount until it worked with 1352 data point and increase sleeping time to 10 sec.

Conclusion & future work

Data obtained via scraping websites are not reliable enough to validate a model's performance and predict accurate results. The features were not enough, and the data points were relatively few

In future work we will do more effort to obtain more related features to the target such as book field. And we will spend more time to increase data amount and arise degree of freedom.

