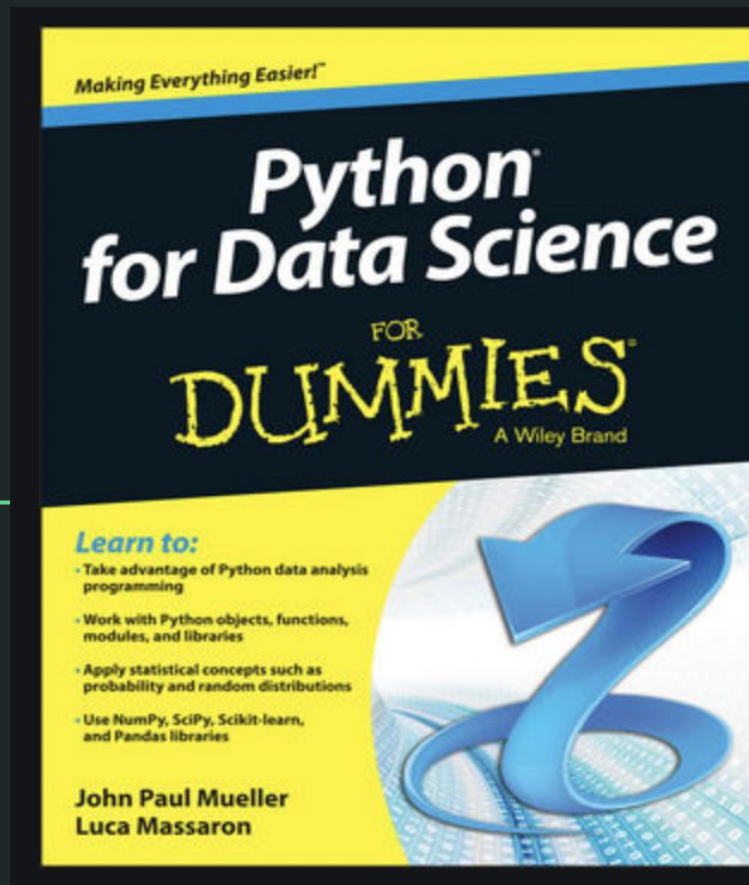


Judging a Book by... its Characteristics!

Linear Regression / Web Scraping

Metis Bootcamp
Project 2
Brandon McNeil





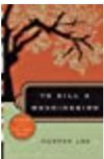

Predicting User Rating from GoodReads.com:

Context:

Goodreads.com is a free online platform where users can rate books on a scale of 1 to 5.

Our Analysis Goal:

Can we predict a book's average user rating on **Goodreads.com** based on the book's characteristics?

		All Votes	Ac
1		The Hunger Games (The Hunger Games, #1) by Suzanne Collins ★★★★★ 4.32 avg rating — 6,689,667 ratings <input type="button" value="Vote For This Book"/> score: 3,072,099, and 31,307 people voted	
2		Harry Potter and the Order of the Phoenix (Harry Potter, #5) by J.K. Rowling ★★★★★ 4.50 avg rating — 2,655,213 ratings <input type="button" value="Vote For This Book"/> score: 2,698,437, and 27,595 people voted	
3		To Kill a Mockingbird by Harper Lee ★★★★★ 4.28 avg rating — 4,751,816 ratings <input type="button" value="Vote For This Book"/> score: 2,334,446, and 23,986 people voted	
4		Pride and Prejudice by Jane Austen ★★★★★ 4.27 avg rating — 3,188,554 ratings <input type="button" value="Vote For This Book"/> score: 2,071,427, and 21,344 people voted	

Data Collection via Web Scrapping:

goodreads Home My Books Browse Community Search books

A Dance with Dragons
(A Song of Ice and Fire #5)
by George R.R. Martin

★★★★★ 4.32 Rating details 576,103 ratings 22,532 reviews

Alternate cover edition of ASIN B004XISI4A

In the aftermath of a colossal battle, the future of the Seven Kingdoms hangs in the balance—beset by newly emerging threats from every direction. In the east, Daenerys Targaryen, the last scion of House Targaryen, rules with her three dragons as queen of a city built on dust and death. But Daenerys has thousands of enemies, and ma ...more

Want to Read
Rate this book
★★★★★

GET A COPY
Amazon Stores

Kindle Edition, 1125 pages
Published July 12th 2011 by Bantam

Original Title A Dance with Dragons
Edition Language English
Series A Song of Ice and Fire #5
Characters Brandon Stark, Tyrion Lannister, Daenerys Targaryen, Theon Greyjoy, Arya Stark...more
Literary Awards Hugo Award Nominee for Best Novel (2012), Locus Award for Best Fantasy Novel (2012), World Fantasy Award Nominee for Best Novel (2012), SFX Award for Best Novel (2012), British Fantasy Award Nominee for Best Novel (2012) ...more

Other Editions (175)

...Less Detail

Share
Recommend It | Stats | Recent Status Updates

READERS ALSO ENJOYED

See similar books...

GENRES

Fantasy	23,078 users
Fiction	4,767 users
Fantasy > Epic Fantasy	1,035 users
Science Fiction Fantasy	765 users

Data Features Scrapped:

- 1) Title
- 2) Rating
- 3) Num of Ratings (target feature)
- 4) Num of Written Reviews
- 5) Book Type
- 6) Num of Pages
- 7) Published Language
- 8) Is it a Series?
- 9) Character Count
- 10) Literary Awards
- 11) First Listed Genre

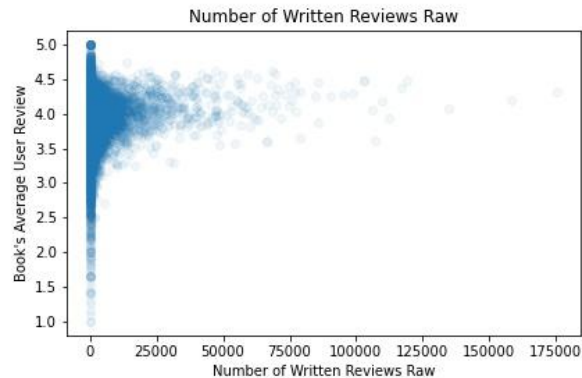
13,000 + records collected!.

Feature Engineering:

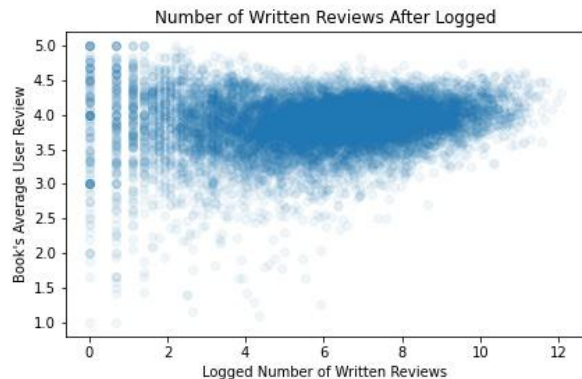
Important Callouts:

- Removed outliers from data that skewed our data (ex. books with over 2000).
- Logged **Number of Written Reviews** to fix distribution issues.
- Created Dummy variables for a **Book's listed Genre** which created around 50 new variables.

Before Log Function



After Log Function



Cross Validation:

- Initial testing indicated Simple Linear Regression and Ridge Regression performed the best on data.
- After running optimal models through KFold cross-validation, both received the same scores - so we kept things simple!

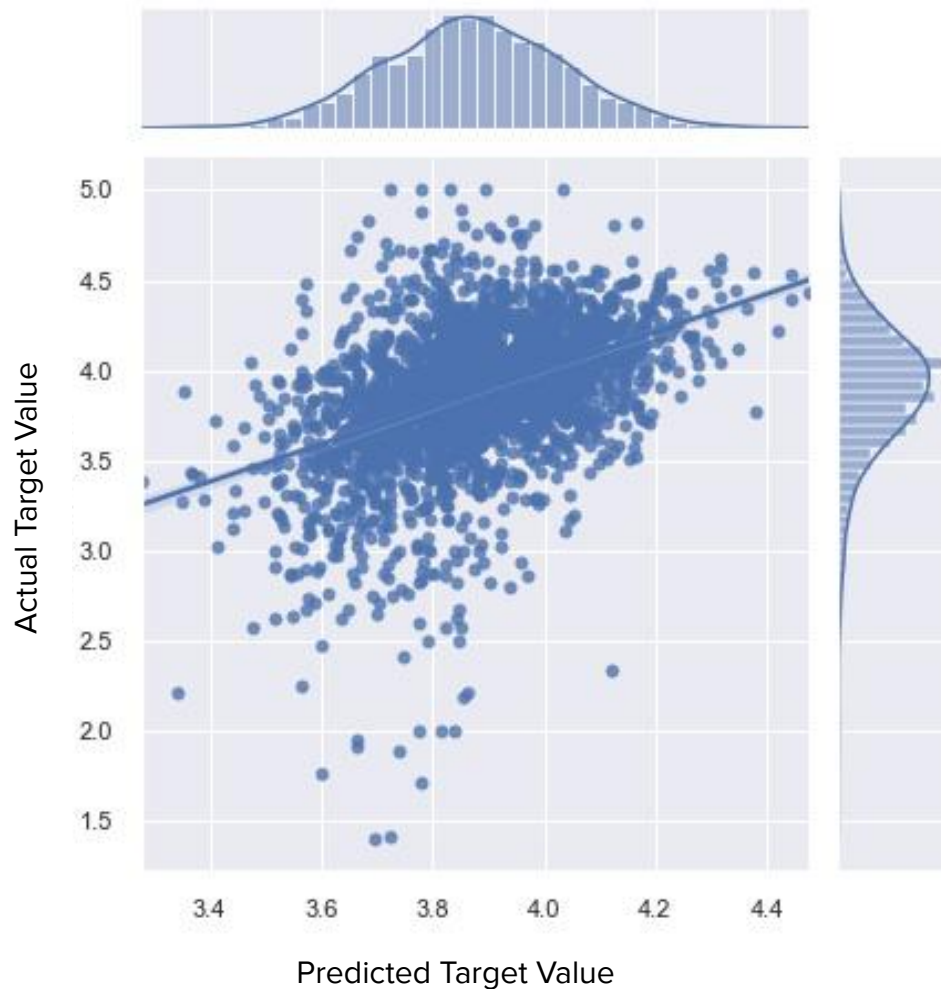
Simple mean cv r^2 : 0.1725 +- 0.0128

Ridge mean cv r^2 : 0.1725 +- 0.0128

Results - Predicted vs Actual:

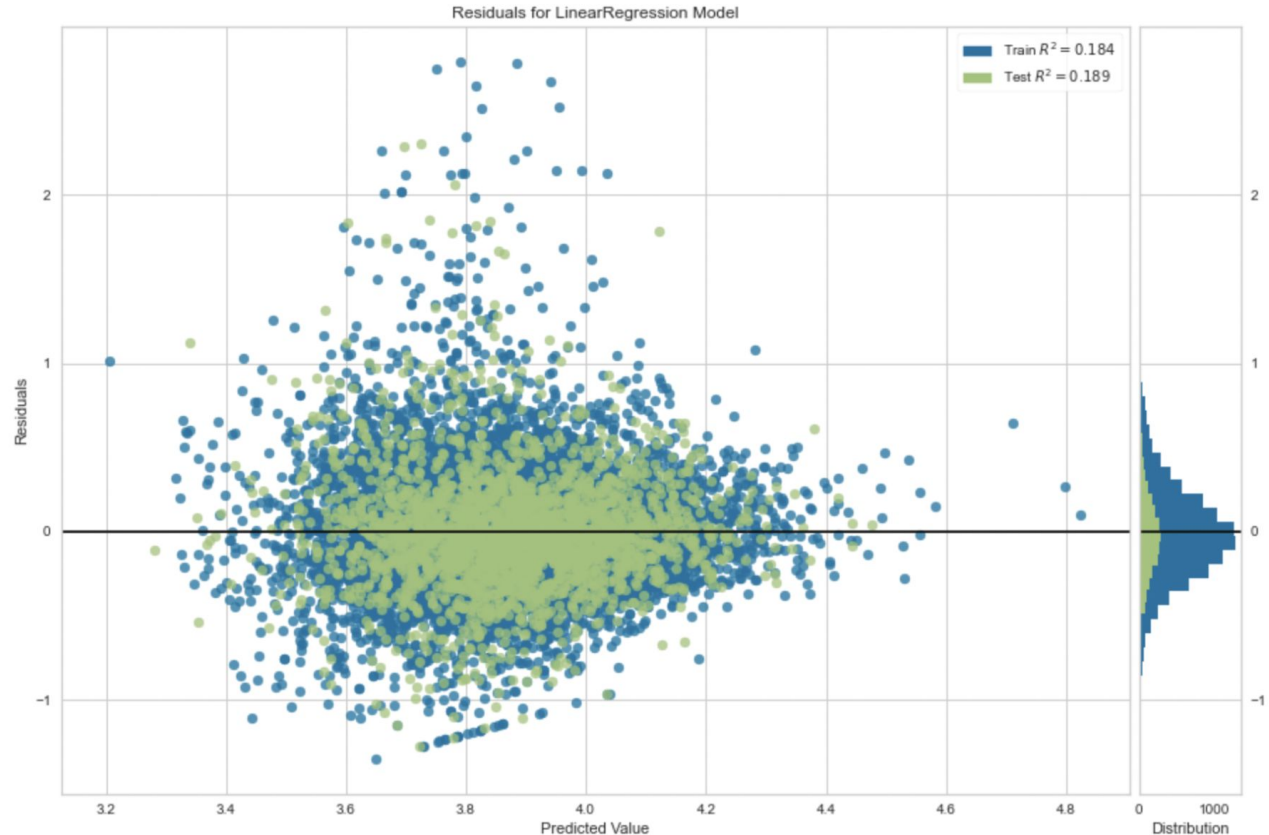
Final R-Squared: 0.184

Predicted and Actual targets are not symmetrical around the line & do not align neatly to it.



Results: Residual Plotting

- While there are some floating points above 2 on the y-axis, most of the clusters are forming around 0-1
- Not a fully symmetrical distribution so improvement is needed!



Key Takeaways:

Predicting reviews based on book features is difficult!

Reviews are a subjective and therefore more elements are required to capture that subjectivity!

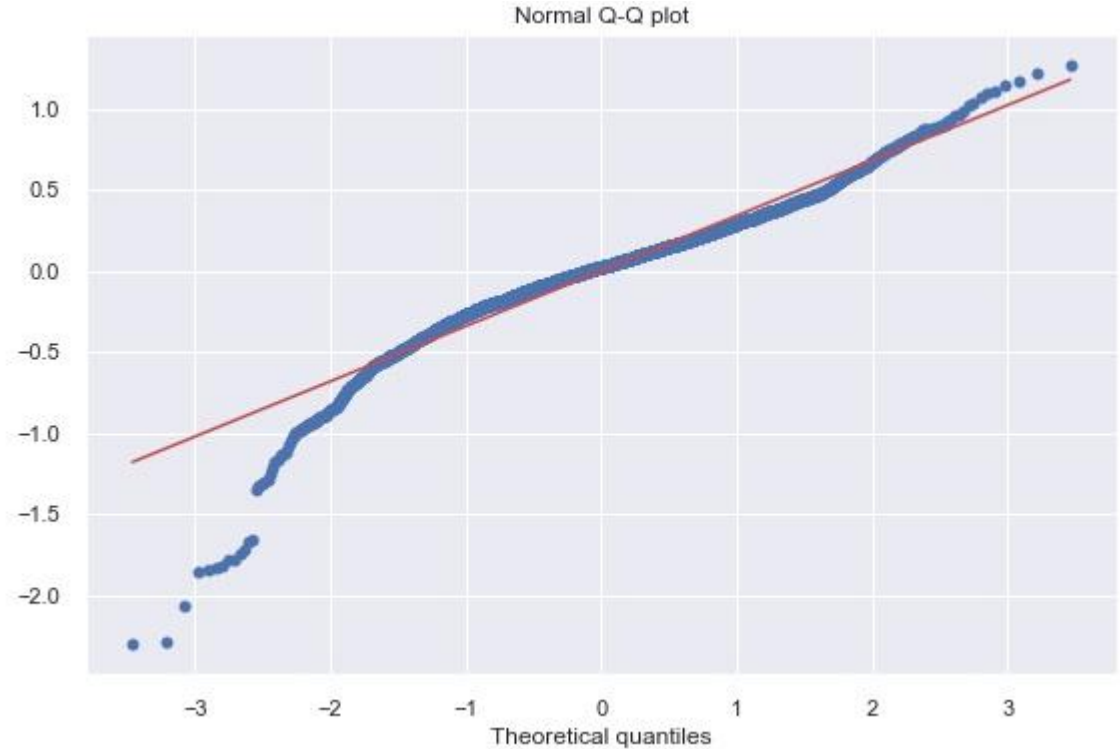
Future iterations of this project will look to analyze user's written reviews to aggregate common words to see if we can peel back the subjective blocker.

Appendix - Additional Results:

1. Sum of Squared Error (SSE) - 317.564
2. Mean Absolute Error (MAE) - 0.249
3. Root Mean Squared Error (RMSE) - 0.352

Appendix - Q-Q Plot of Residuals:

Our residuals do not fit the line as well as we'd like!



Appendix - Final Coefficients:

```
( 'book_cover_type', 0.00035134685747900376),  
( 'num_of_pages', -0.09070320517401742),  
( 'foreign_language', 0.12298710079757218),  
( 'is_series', 0.05619773788479662),  
( 'won_awards', 0.007462186959562189),  
( 'num_of_chars', 0.022623215338221148),  
( 'genre_1', -0.14942867531347964),  
( 'num_written_reviews', -0.11379753004007456) ]
```