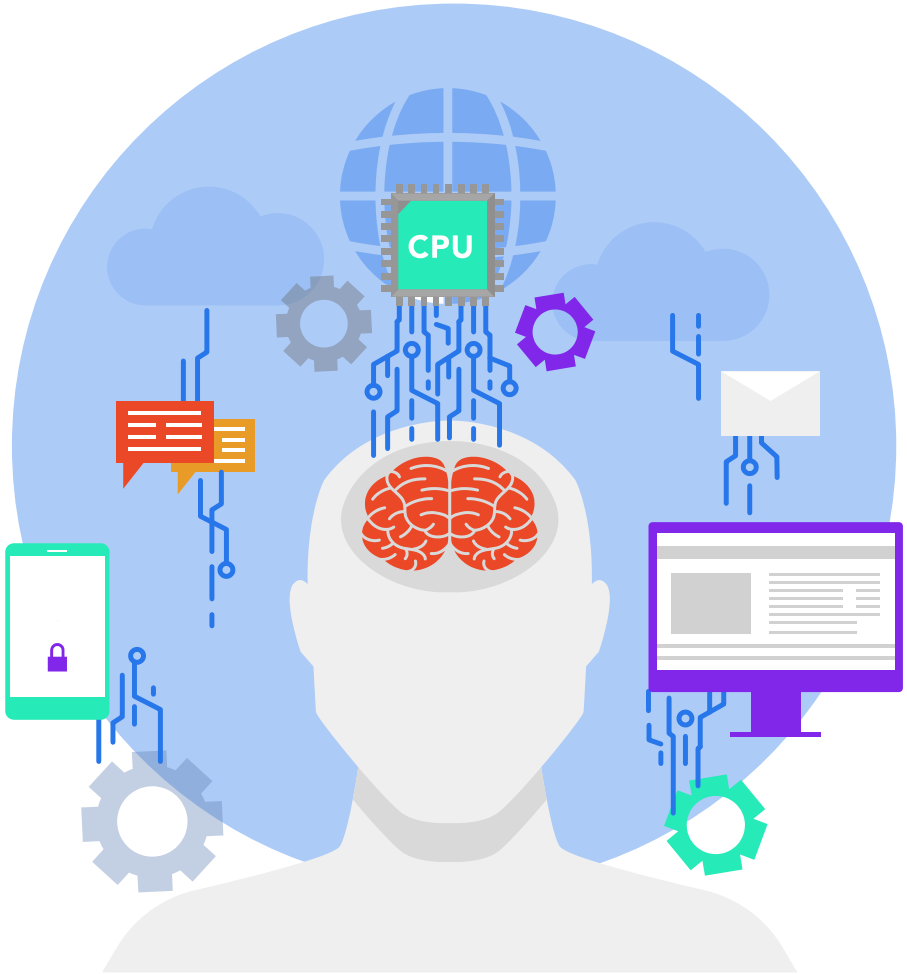


Machine Learning & Data Analysis

Ali Haider
Giacomo Pedemonte

Dataset: Amazon Top 50 Bestselling Books 2009 – 2022

In this dataset there are 7 features that represent the most important characteristics for bestsellers.

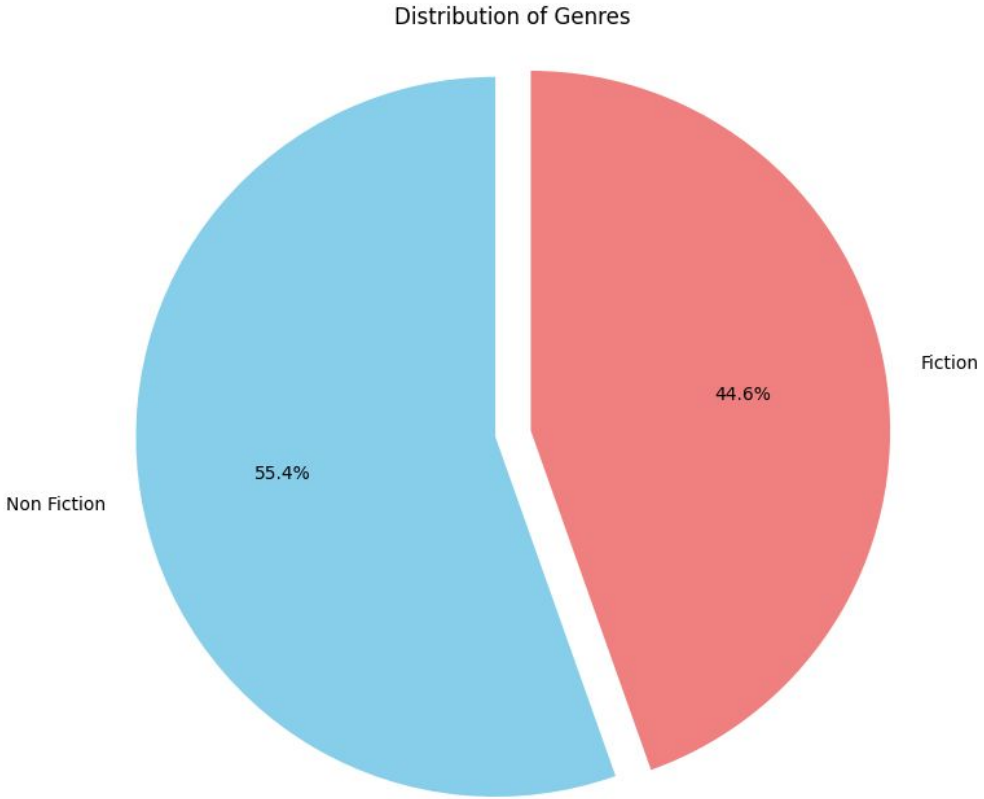
	Name	Author	User Rating	Reviews	Price	Year	Genre
0	Act Like a Lady, Think Like a Man: What Men Re...	Steve Harvey	4.6	5013	17	2009	Non Fiction
1	Arguing with Idiots: How to Stop Small Minds a...	Glenn Beck	4.6	798	5	2009	Non Fiction
2	Breaking Dawn (The Twilight Saga, Book 4)	Stephenie Meyer	4.6	9769	13	2009	Fiction
3	Crazy Love: Overwhelmed by a Relentless God	Francis Chan	4.7	1542	14	2009	Non Fiction
4	Dead And Gone: A Sookie Stackhouse Novel (Sook...	Charlaine Harris	4.6	1541	4	2009	Fiction
...	...	...	...	...	...	...	...
695	The Wonderful Things You Will Be	Emily Winfield Martin	4.9	20920	9	2022	Fiction
696	Ugly Love: A Novel	Colleen Hoover	4.7	33929	10	2022	Fiction
697	Verity	Colleen Hoover	4.6	71826	11	2022	Fiction
698	What to Expect When You're Expecting	Heidi Murkoff	4.8	27052	13	2022	Non Fiction
699	Where the Crawdads Sing	Delia Owens	4.8	208917	10	2022	Fiction

The Dataset contains

- 700 Books
- 441 Unique Titles
- 305 Authors

The dataset doesn’t contain empty values and there is no cleaning to make on the dataset.

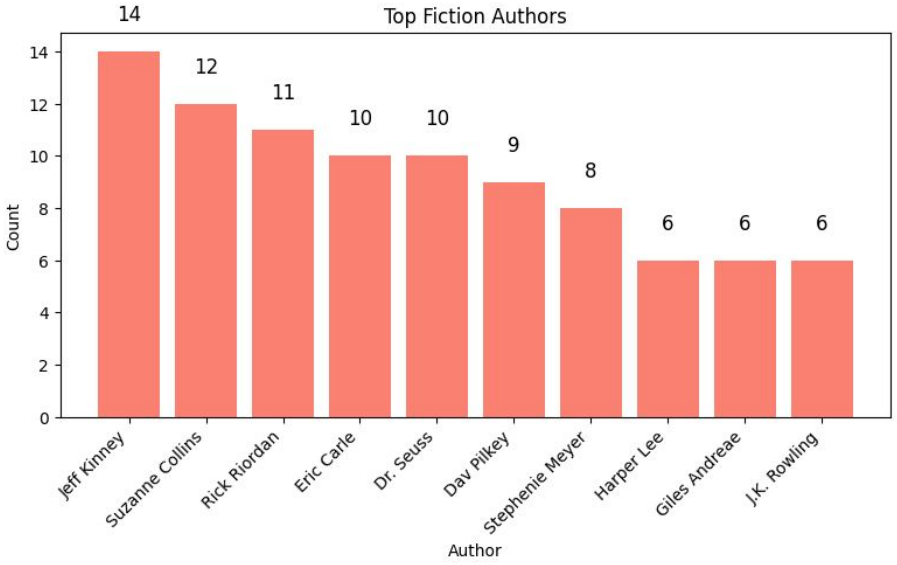
Dataset: Amazon Top 50 Bestselling Books 2009 – 2022

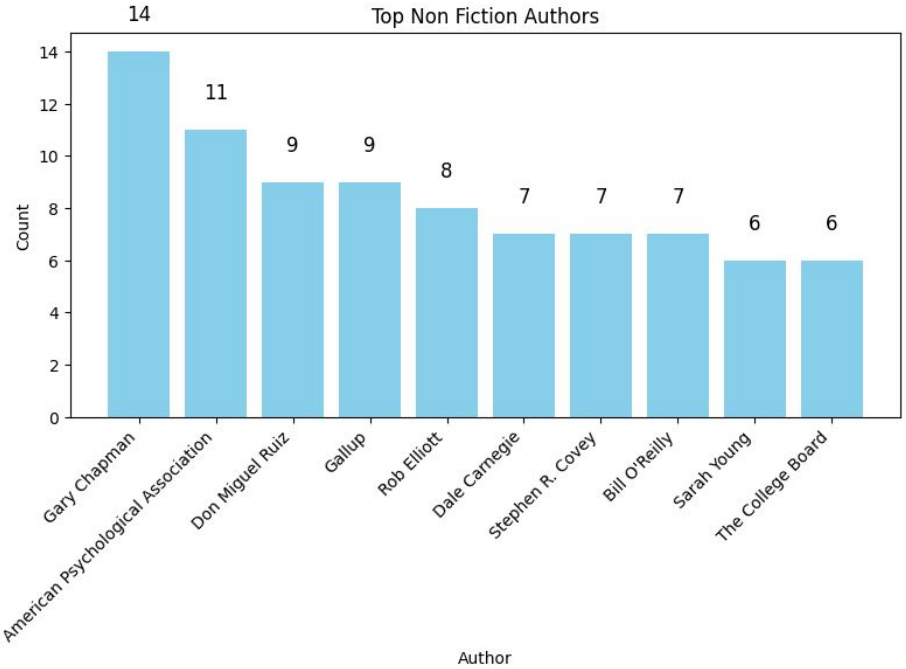


Genres: Fiction & Non Fiction

It seems that a shift towards Non-Fiction may be beneficial for writers

Data Analysis - Authors





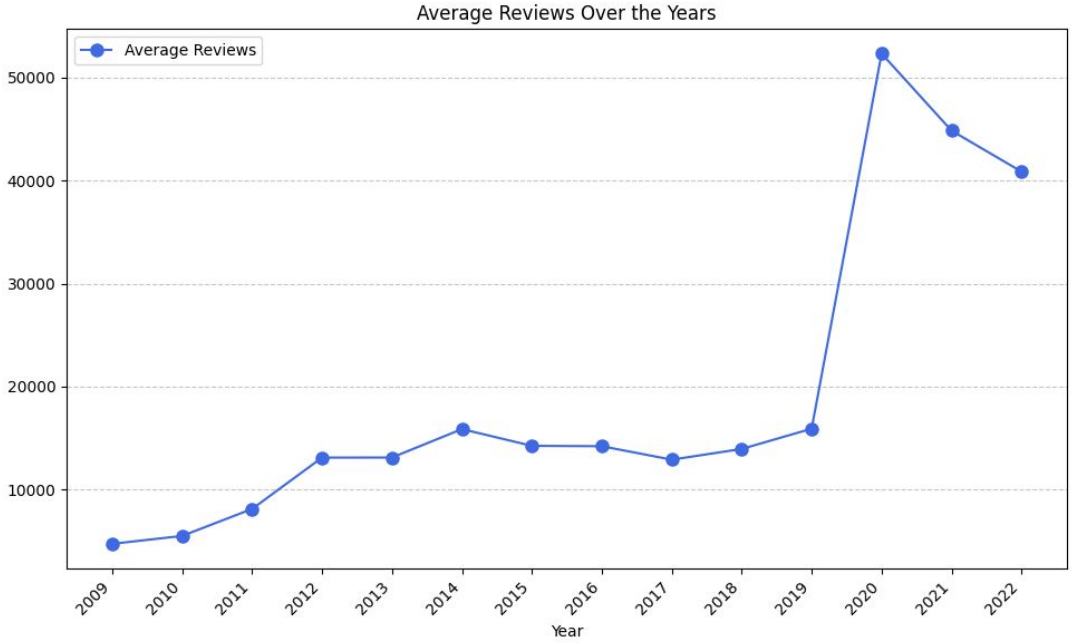
Jeff Kinney and Gary Chapman are respectively the most present artist in the two Genres

Data Analysis - Reviews

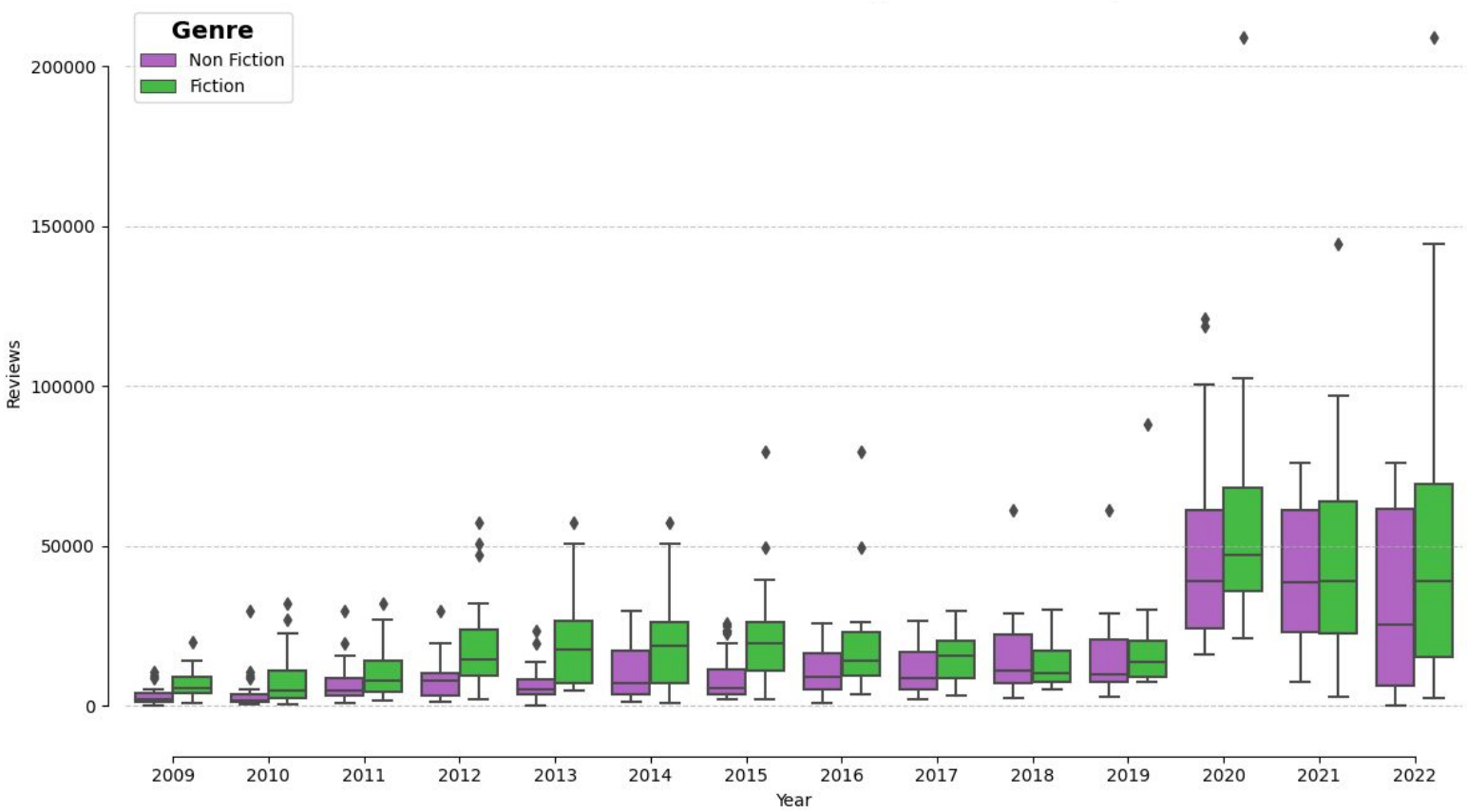
Noticeable Increase in User Reviews for Bestsellers Over the Years

2020 Shows a Significant Spike in User Reviews

Possibly Linked to the COVID-19 Pandemic Encouraging Reading Habits



Data Analysis - Reviews by Genre over the Years



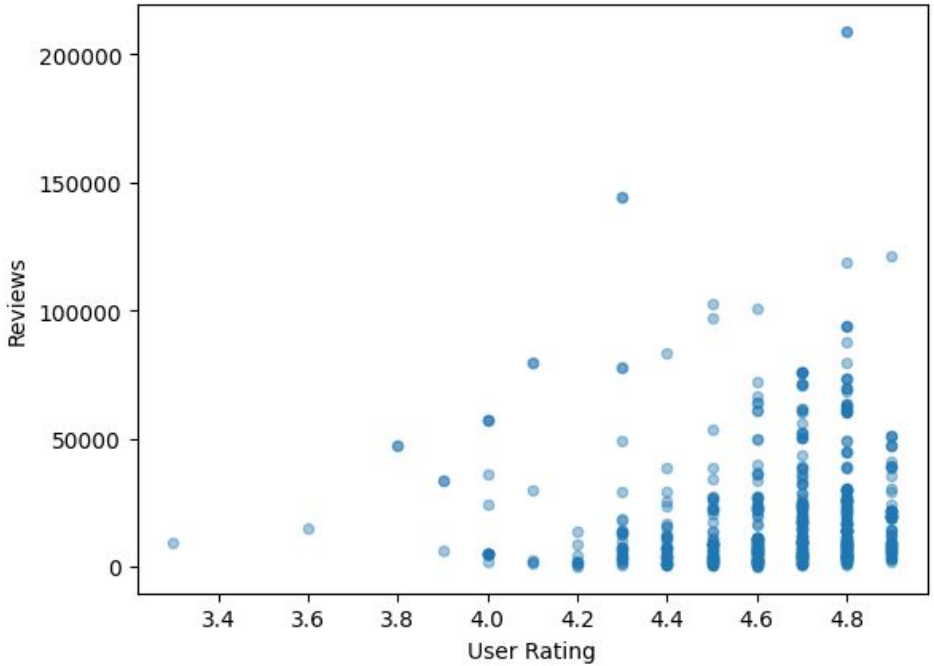
Data Analysis - User Rating and Reviews Analysis

Books with relatively low Ratings have fewer Reviews

No Observable Correlation Between User Rating and the Number of Reviews

A high User Rating doesn't Necessarily Mean more Reviews, and vice versa.

So let's combine normalized reviews and ratings to calculate book popularity.



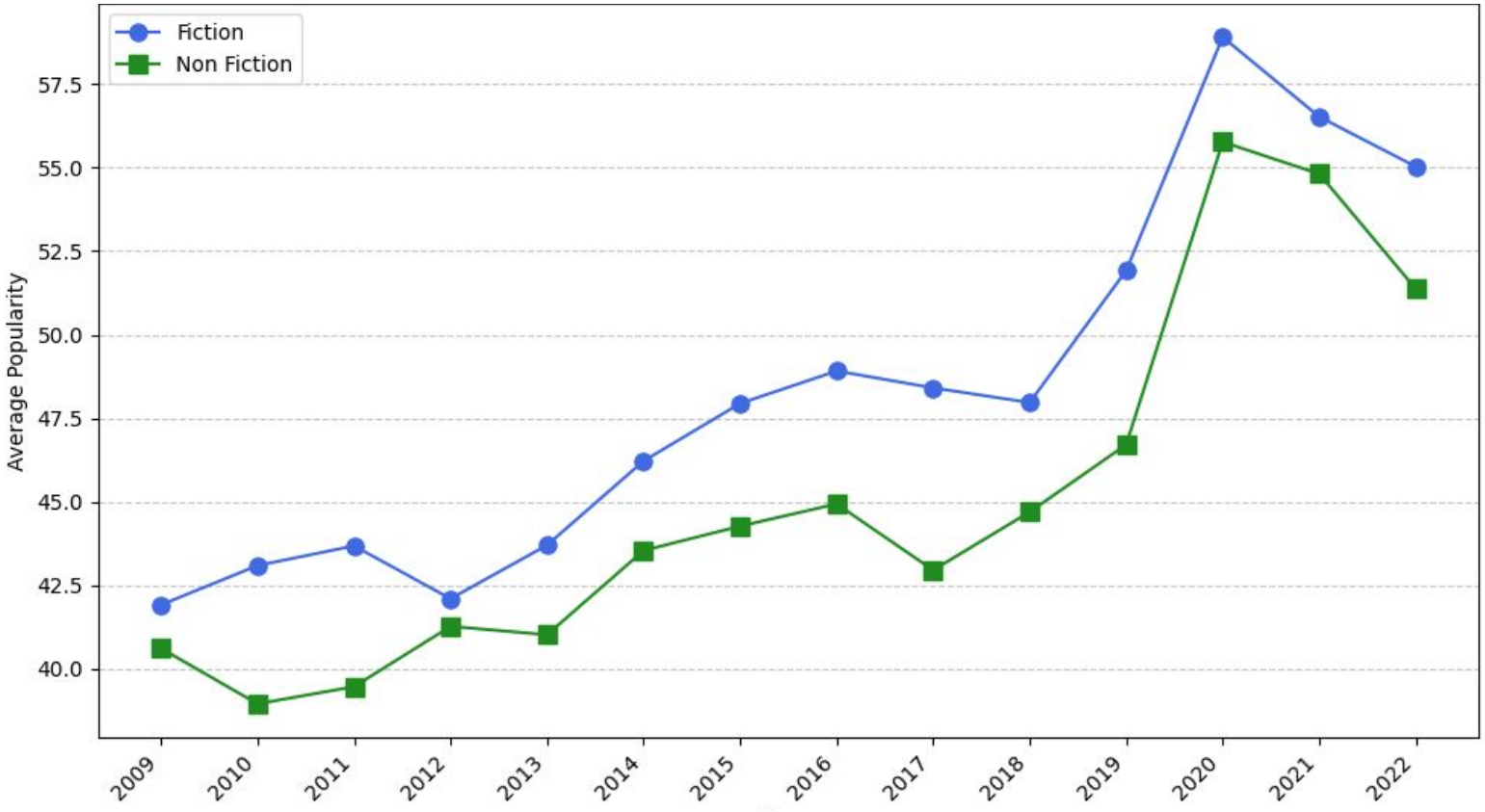
Data Analysis - Popularity

Name	Author	User Rating	Reviews	Price	Year	Genre	Reviews_Normalized	UserRating_Normalized	Price_Normalized	Popularity
Where the Crawdads Sing	Delia Owens	4.8	208917	10	2022	Fiction	100.000000	93.75	9.523810	96.875000
Where the Crawdads Sing	Delia Owens	4.8	208915	12	2020	Fiction	99.999043	93.75	11.428571	96.874521
A Promised Land	Barack Obama	4.9	121109	16	2020	Non Fiction	57.962466	100.00	15.238095	78.981233
Becoming	Michelle Obama	4.8	118767	21	2020	Non Fiction	56.841249	93.75	20.000000	75.295624
The Boy, the Mole, the Fox and the Horse	Charlie Mackesy	4.8	93749	10	2022	Fiction	44.864037	93.75	9.523810	69.307018

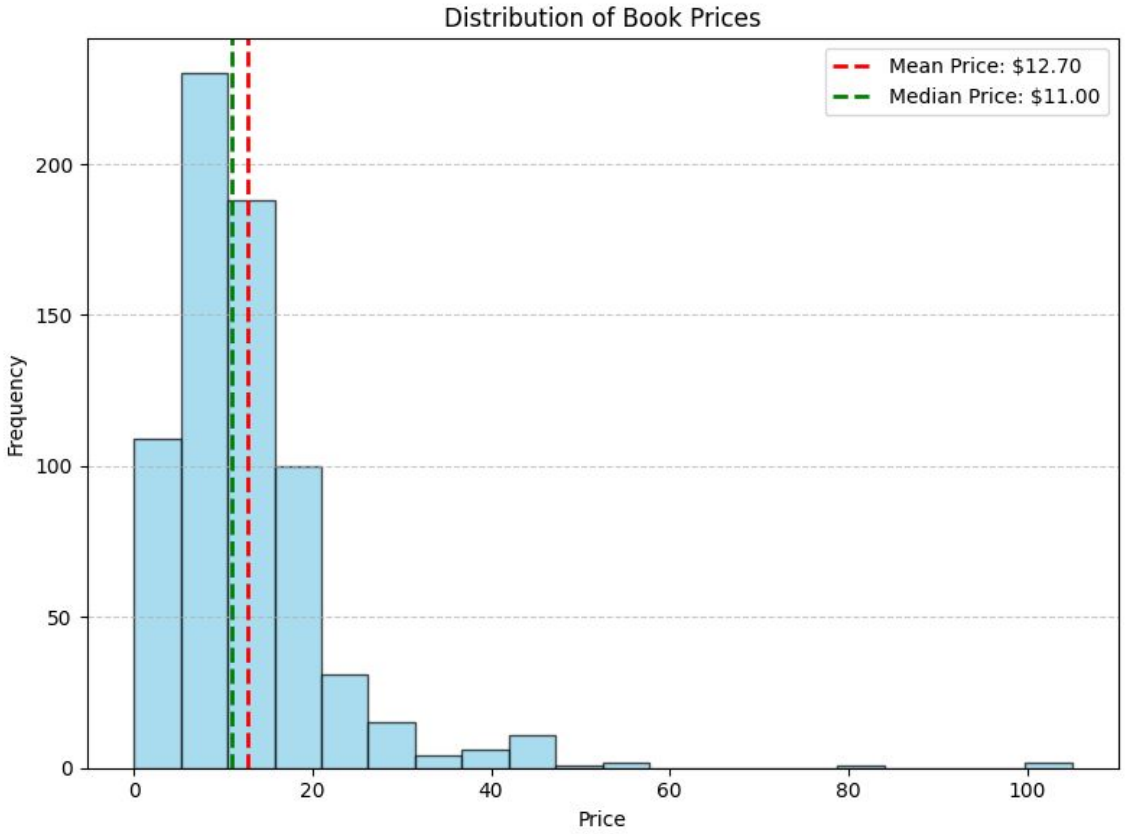
We combined normalized reviews and ratings to calculate book popularity.

This approach provides a more comprehensive view of book appreciation

Data Analysis - Popularity by Genre Over the Years



Data Analysis - Price Over the Years



Most Bestsellers priced below $20

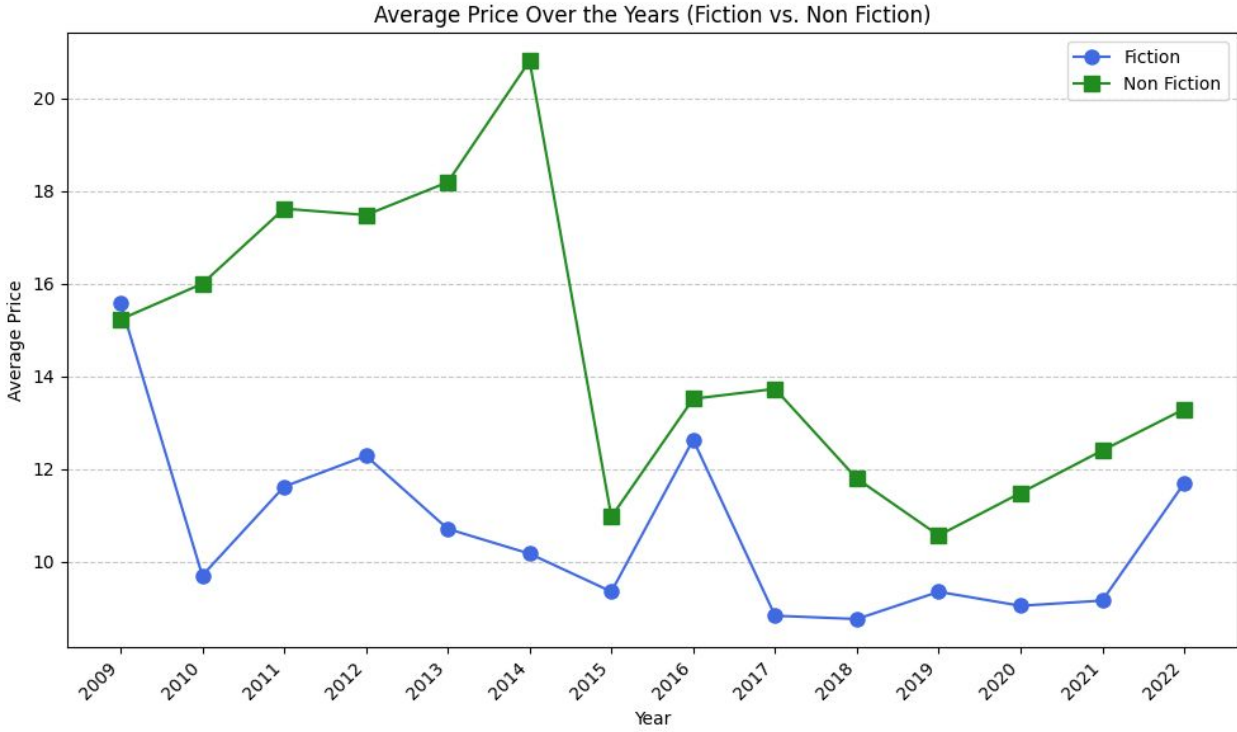
We can say that affordable books attract new and Non-Traditional readers

Data Analysis - Price by Genre Over the Years

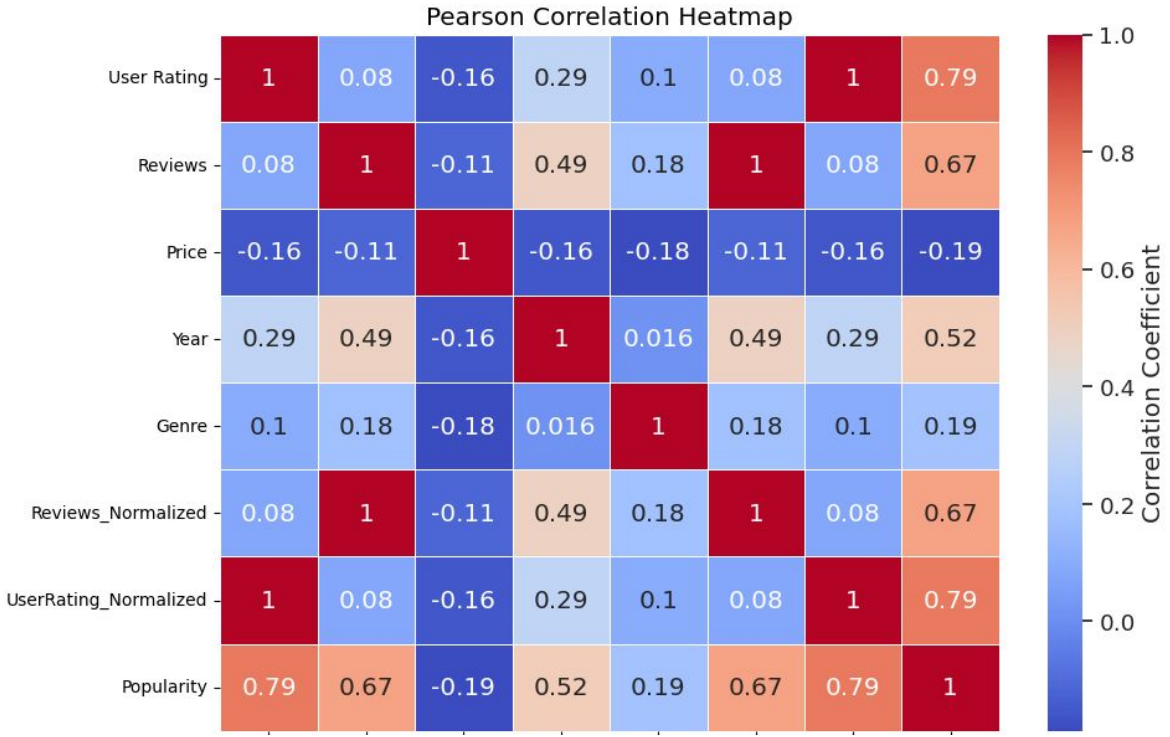
Non-Fiction books have consistently commanded higher prices on average.

This is because Non-Fiction authors aim to maximize their profits, often targeting readers willing to invest in knowledge.

In contrast, Fiction authors may price their books more affordably as they work to build a dedicated reader base for future works.



Data Analysis - Correlation of the Features



There are no features that are correlated with each other.

The only features that seems to be correlated are the Year and the Popularity as we have discovered also in previous analysis: in the last years the popularity of books increases.

Machine Learning

Given the absence of strong feature correlations, we adjusted our analysis approach. We proactively explored correlations in other areas of the dataset.

This adaptability allowed us to uncover valuable insights:

- Genre Prediction using the books name
- Books Recommendation System using text features
- Price Prediction using all the books features

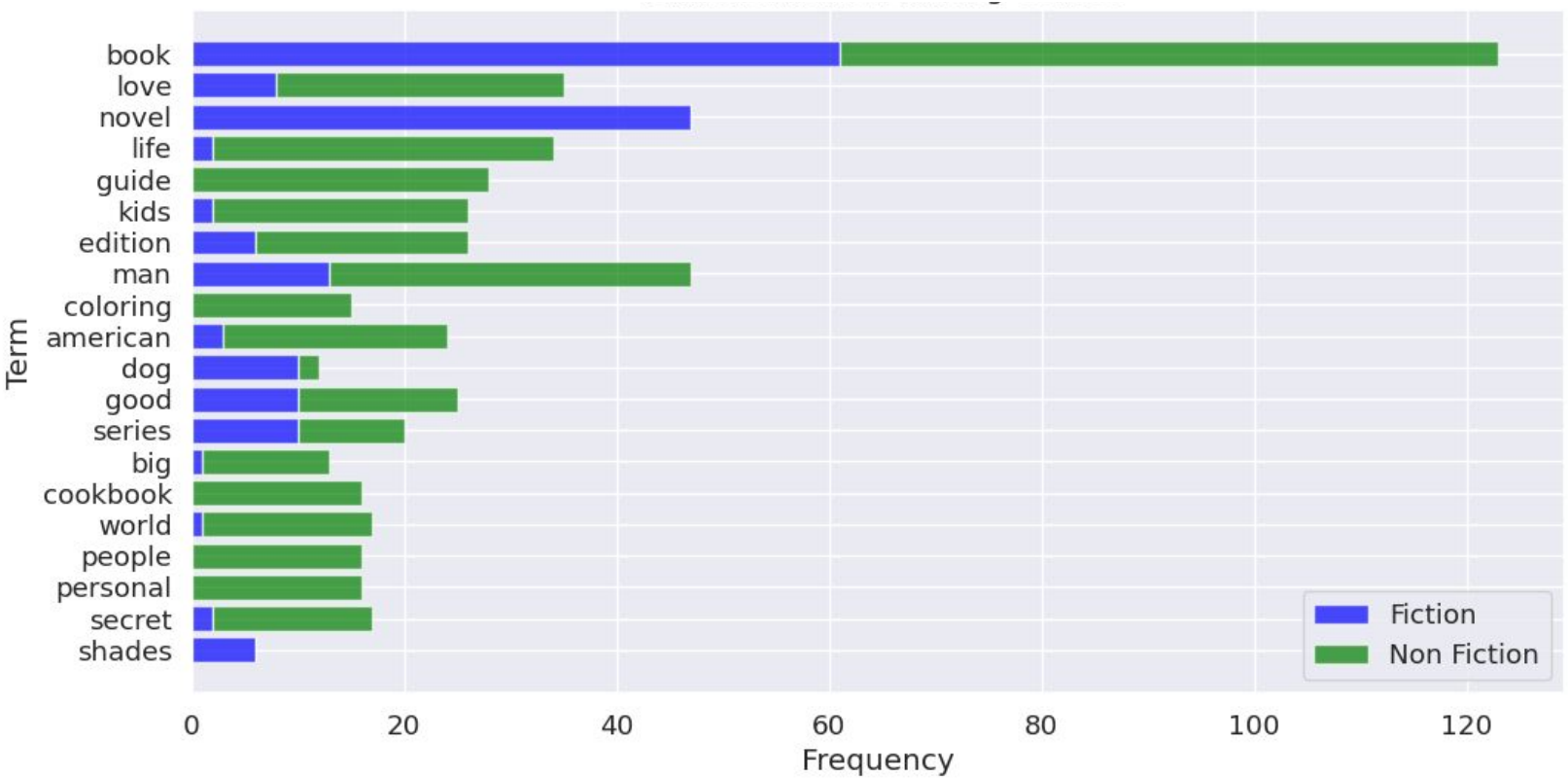
Machine Learning - Genre Prediction

To make the prediction of the Genre easier we assign a numerical representation to the Genre feature:
Non Fiction = 0, Fiction = 1

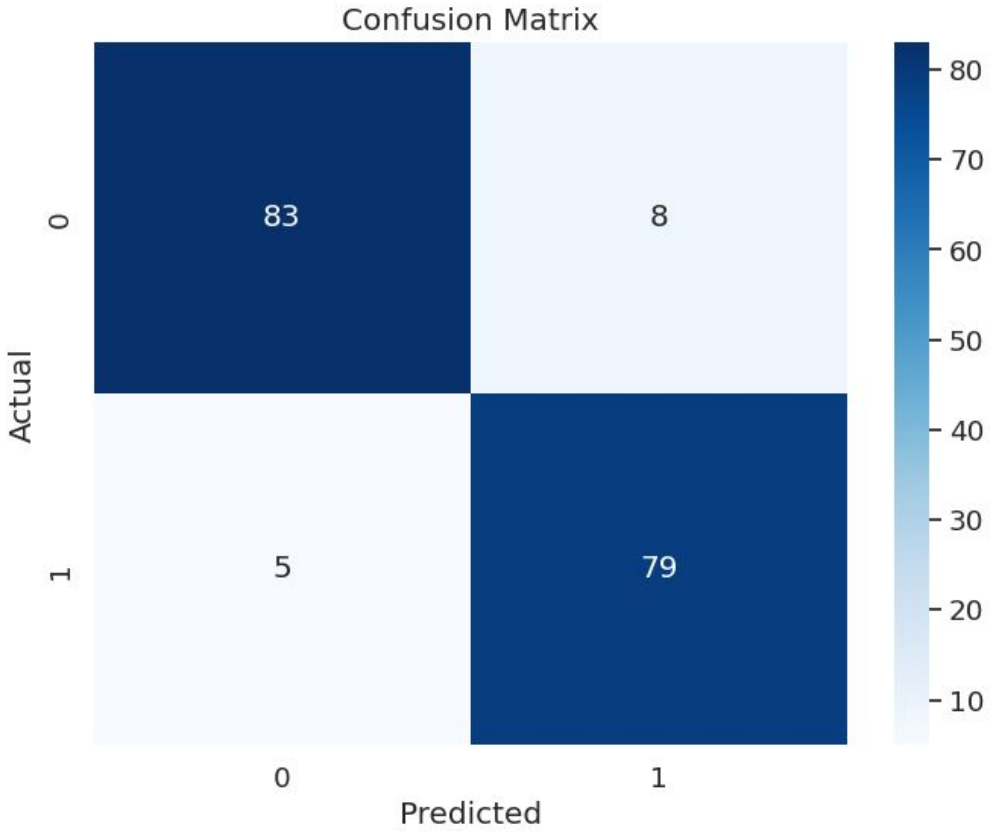
X = Name
y = Genre

- Counting words in texts, excluding stop words (such as 'a,' 'an,' 'the,' ...)
- Transforming the texts into a Document Term Matrix (DTM): this will enable us to analyze the frequency of key words in the texts
- Data Split for Training and Validation
- Training a Logistic Regression model on our Training set
- Evaluating the Model on Validation Data

Machine Learning - Genre Prediction
Terms Distribution



Machine Learning - Genre Prediction Evaluation

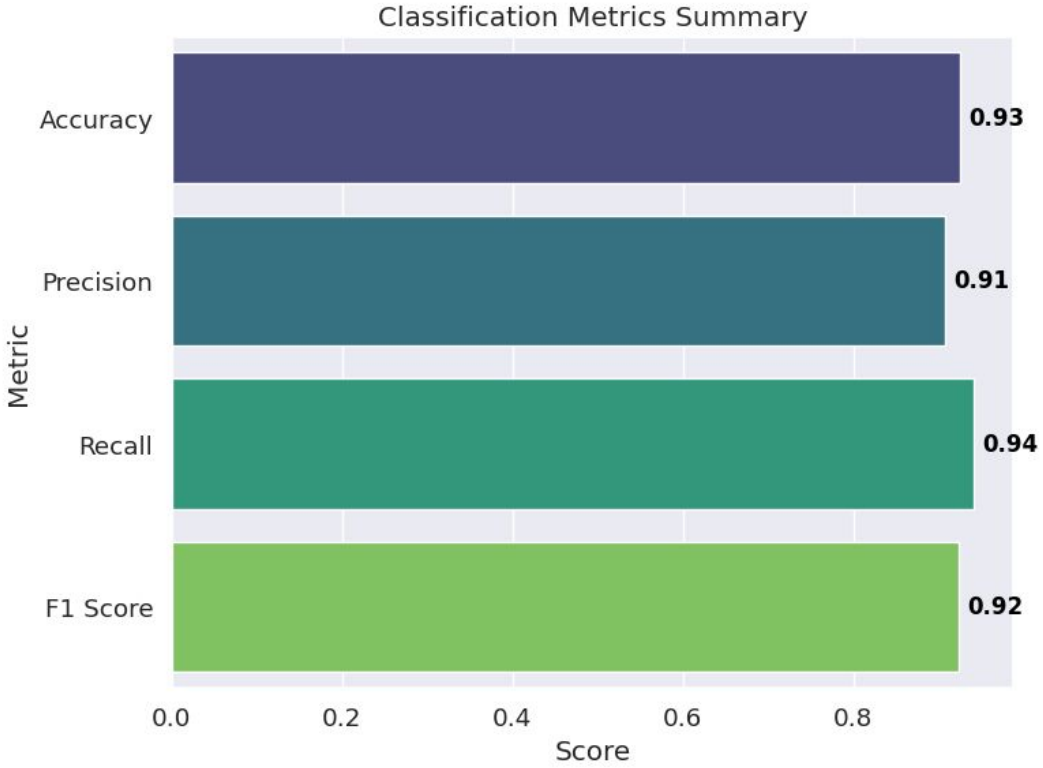


To evaluate our Model we've used the Confusion Matrix. Since it is a binary classification problem this matrix have dimension 2*2.

The top left square represents the right prediction of the Non Fiction books by checking the y_test with the y_pred *(TN)*.

The bottom right square represents the right prediction for Fiction books *(TP)*.

Machine Learning - Genre Prediction Evaluation



Analyzing those metrics we can assume that our model predict with high precision and accuracy the genre of the book with only the name as input.

This is confirmed also by the F1 score which is the harmonic mean of the Precision and Recall metrics.

The Recall metrics indicates us the number of TP(Fiction books) over the total actual number of the Fiction Books.

Machine Learning - Book Recommendation

We want to recommend books similar to the book's title that the user insert as input.

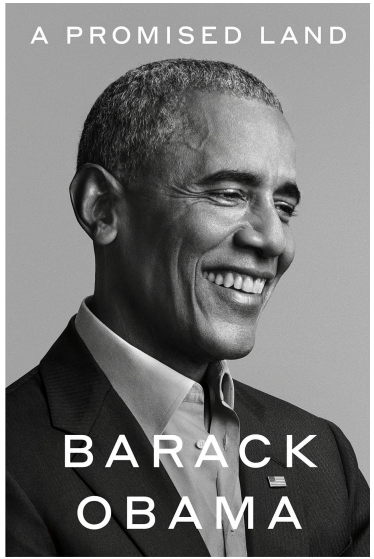
- First of all we agglomerate the three features that we will use for the Content-Based Filtering:
  - Name
  - Author
  - Genre

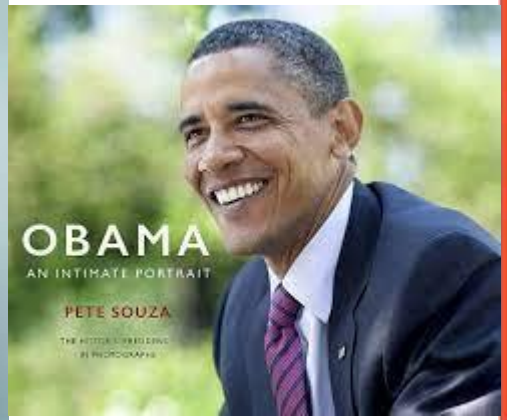
- This combination will be transformed in a TF-IDF(Term Frequency-Inverse Document Frequency) vector for text features, which helps in quantifying the importance of words in the text.

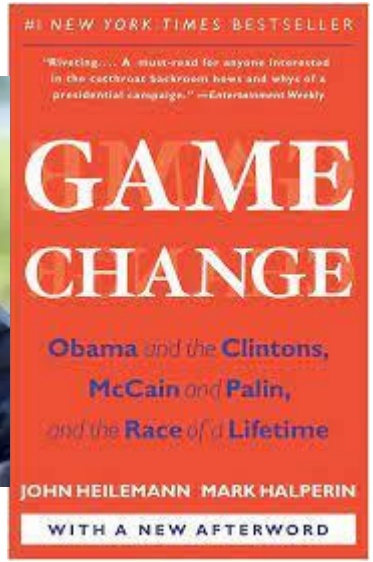
- Then calculates the cosine similarities between books based on their text features using the linear kernel method, which can be useful for tasks like recommendation systems or text-based similarity analysis.

Let's see how it works!

Machine Learning - Book Recommendation







Input:
A promised Land

Output:
Becoming, Michelle Obama
Obama: An Intimate Portrait
Game Change: Obama and the Clintons, McCain and Palin, and the Race of a Lifetime

Machine Learning - Book Recommendation

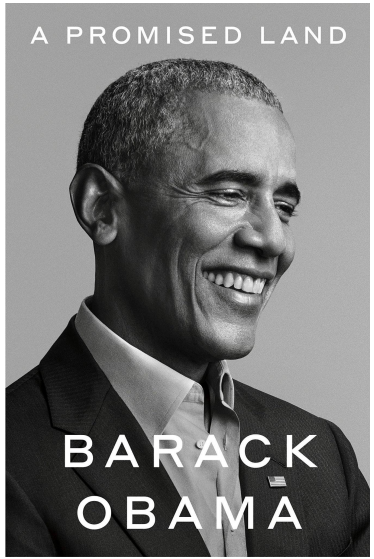
To understand if the above Content-Based Filtering Model was the best to obtain this book recommendation, we try with different Model called Hybrid-Approach that makes us consider also the quantitative features.

We are now combining the TF-IDF Matrix with the normalized values of the quantitative data inside the dataset like:

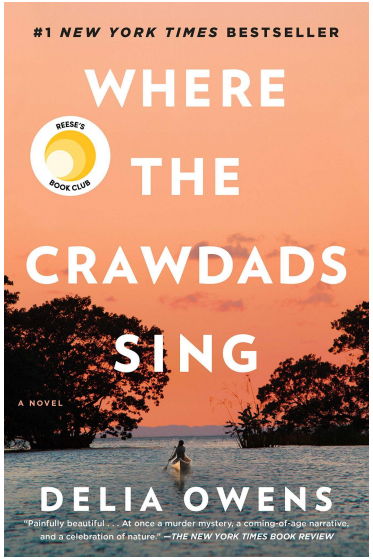
- User Rating
- Reviews
- Popularity.

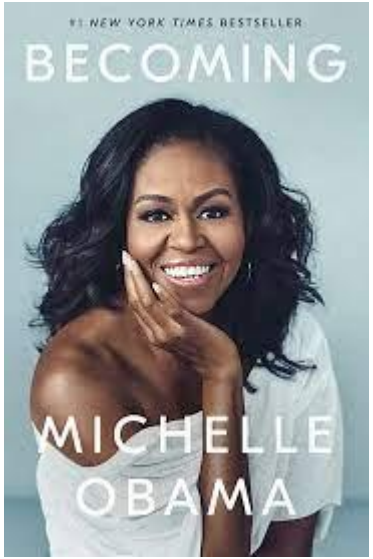
As it has emerged from the previous analysis the low correlation between the feature will makes us obtain a less precise result corresponding on the Content-Based Filtering.

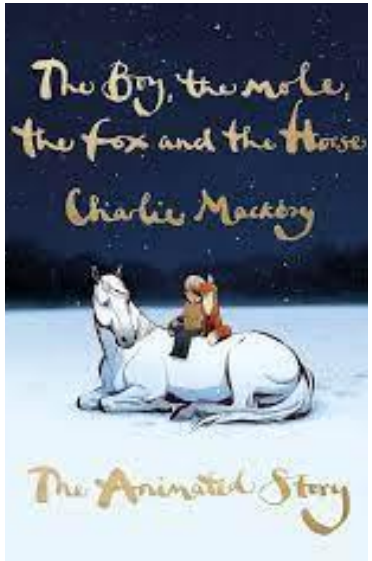
Machine Learning - Book Recommendation



Input:
A promised Land







Output:
Where The Crawdads Sing
Becoming, Michelle Obama
The Boy, the Mole, the Fox and the Horse

Worst Results!

Machine Learning - Price Prediction

Until now, we doesn't used well the numerical features to obtain prediction because we have seen in the analysis that there aren't feature correlated to each other.

Now we want to understand if it is possible to predict the price using all the features of the books.

So we consider to make the prediction of the book's price:

- Name
- Author
- User Ratings
- Reviews
- Year
- Popularity

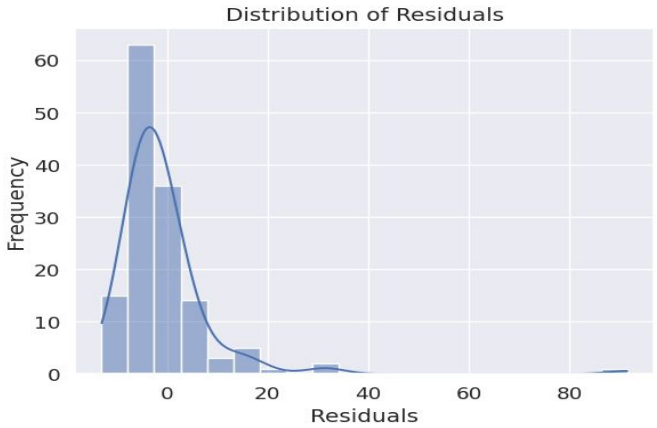
We start using a Ridge Regression then applied Kernels with Kernel-Ridge Regression and make a Cross-Validation to tune the hyperparameters to find the best alpha and gamma that makes the regression fit better the data and obtain more precision of the model.

Machine Learning - Price Prediction
Ridge

Evaluation	Value
Mean Squared Error (MSE)	109.76
Root Mean Squared Error (RMSE)	10.48
Mean Absolute Error (MAE)	5.84
R-squared (R2)	0.02

Since the Errors are very high we cannot consider this model good to predict the price.
The value of alpha used was alpha = 0.001

We can try with Kernels to obtain a better result.

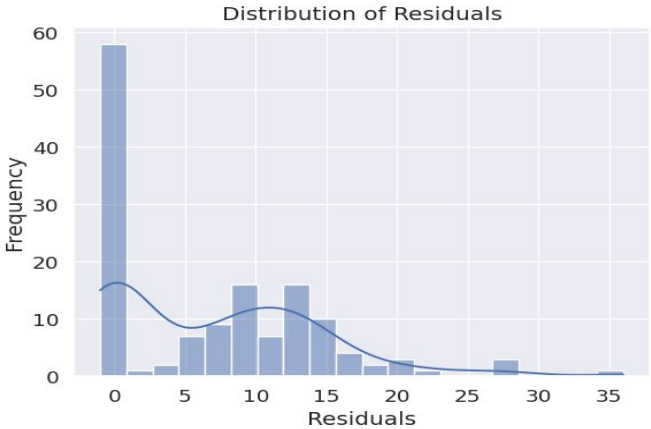


Machine Learning - Price Prediction
KernelRidge

Evaluation	Value
Mean Squared Error (MSE)	102.65
Root Mean Squared Error (RMSE)	10.13
Mean Absolute Error (MAE)	7.00
R-squared (R2)	0.08

Those value are obtained with hyperparameters, alpha = 0.001 and gamma = 1000.

Since the result is better than before but maybe can we obtain a better result with making Cross-Validation using GridSearch.



Machine Learning - Price Prediction
Cross-Validation

Cross-Validation makes us find the best combination of the hyperparameters: alpha and gamma:

Tuning alpha:

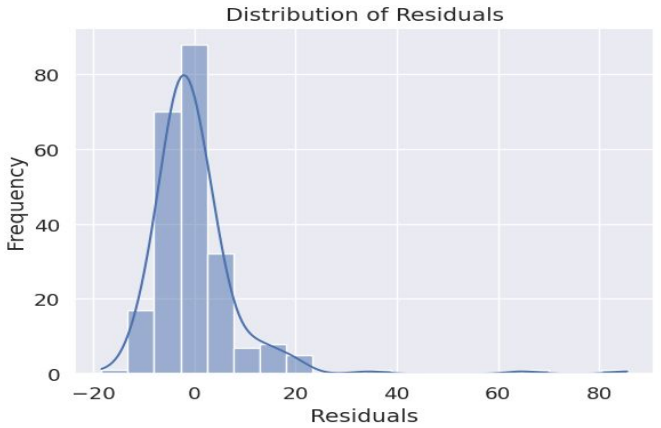
- A smaller alpha allows the model to fit the training data more closely, but it may lead to overfitting.
- A larger alpha imposes stronger regularization, which can help prevent overfitting.
- The optimal value for alpha depends on the dataset and should be tuned to find the best trade-off between bias and variance.

Tuning gamma:

- A smaller gamma value makes us obtain something linear.
- A larger gamma value makes it more peaked.
- Tuning gamma can be important for achieving the right balance between model complexity and generalization.

Machine Learning - Price Prediction
Cross-Validation

Evaluation	Value
Mean Squared Error (MSE)	94.35
Root Mean Squared Error (RMSE)	9.71
Mean Absolute Error (MAE)	5.50
R-squared (R2)	0.20



With the Cross-Validation we obtain as best hyperparameters:
Alpha = 1.0
Gamma = 0.0001

We obtain an improvement in respect of the simple usage of the Ridge Regression(the RMSE now is under 10) but we can’t take this prediction seriously because the Errors remains too high.

Conclusion

To make a quick resume of what we have discover performing the Data Analysis and Machine Learning:

1. Discovered the features inside the datasets analyzing the best genres and finding out the correlation of the features.
2. Machine Learning models to obtain several different prediction:
    a. Prediction of the genre
    b. Book Recommendation System
    c. Prediction of the price
3. The models that makes us obtain a more accurate results were the models that takes consideration of the text features. The poor correlation between the other features makes us obtain not trustable results for the Prediction of the price

THE END

Thanks for the Attention!