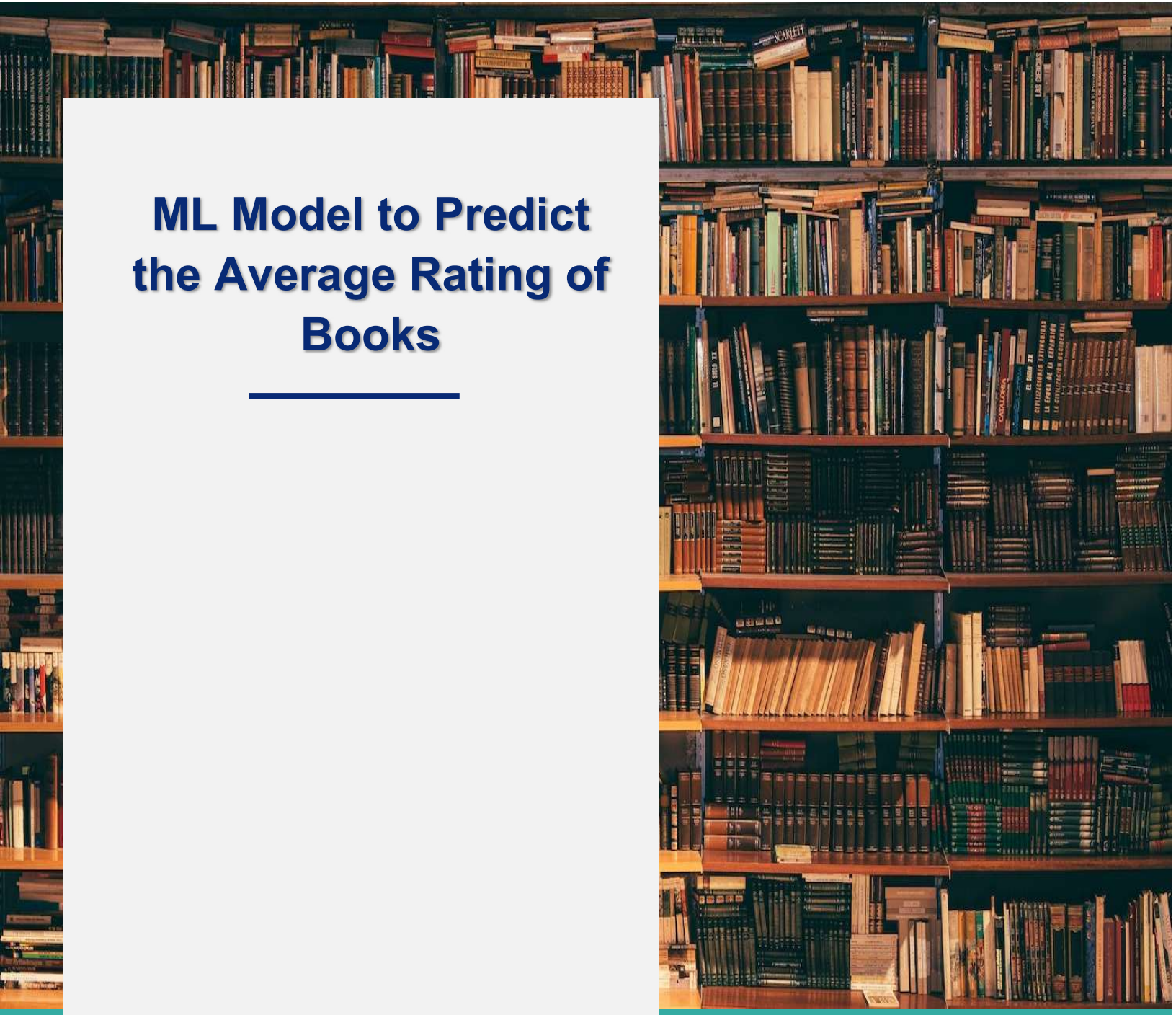


ML Model to Predict the Average Rating of Books

2023-2024

DATA SCIENCE TECH INSTITUTE
Team Members: Ashutosh Yadav, Deepanshu Yadav, Deepanwita Saha, Naresh Kumar

Data ScienceTech Institute

Contents

Summary

Predict the Average Rating of Books

The project is available on GitHub : https://github.com/Ashuayd/Books-Project

In this project, we have worked on books dataset published by Goodreads. The dataset comprised information/data on the following factors: Book ID, Book title, Authors and Co-Authors, Average rating, ISBN and ISBN 13, Language, Number of pages, Number of ratings count and Text reviews count, Publication date and Publisher.

A book lover before purchasing a book may prefer to consider one or more of the above-mentioned factors with average rating, author, publisher being the most important ones.

With the onset of digital revolution, Goodreads as a platform is very popular amongst book readers. They prefer to look at the average rating and the reviews provided by others.

For book readers like us, along with authors, publishers also play an important role. Publishers like Penguin, Harper Collin have always been associated with good books and hence play a very important role in determining the popularity of a book.

The objective behind building the model was to predict the average rating of a book based on various factors which play an important role in the determination.

This model does not judge the books by their cover but by number of pages, language, publisher, and author etc

Let Us Get Started!

Key findings from the data

From the Goodreads books dataset, we seek to find out the following:

- What is average rating for most of the books?
- In which year maximum books were published?
- Does Publication date, number of pages, text reviews count and ratings count really have an impact on the rating of the books?
- Who are the top 20 authors and publishers?
- In which language are most of the books read?

Let us dive into the dataset to gain insights and build a model to predict average rating

Initial Analysis of the Data

From the Goodreads books dataset, we seek to find out the following:

- What is average rating for most of the books?
- In which year maximum books were published?
- Does Publication date, number of pages, text reviews count and ratings count really have an impact on the rating of the books?
- Who are the top 20 authors and publishers?
- In which language are most of the books read?

After analysing the datatype of respective columns, we found that all are categorical variables, except book ID, ratings count and text reviews count.

Then we analysed the summary statistics of only Quantitative/Non-Categorical variables and Qualitative/Categorical values separately.

Upon analysis of Qualitative dataset, we observed that the dataset comprises 11127 rows and no columns have missing data. But the column named co-author has only 4 rows. In the following section, we will deep dive into the co-author column.

	title	authors	average_rating	isbn	isbn13	language_code	num_pages
count	11127	11127	11127	11127	11127	11127	11127
unique	10352	6643	213	11126	11127	31	999
top	The Brothers Karamazov	Stephen King	4.00	3.58	9780439785969	eng	288
freq	9	40	219	2	1	8908	230

Data Cleaning

The dataset comprises 13 columns. But we notice, the 13th column 'Co-Author' has only 4 rows whereas all other columns have 11127 rows. A further examination was carried on.

Upon examining the dataset, we found out that for the 4 rows under discussion, there was a shift of data by 1 column. This means that some of the authors had shifted to the 'Average Rating' column. The average ratings for the respective rows had shifted to the 'ISBN' column and so forth. Hence, we had a 13th column which had publishers' name due to this data shift.

Removing Columns

We removed 'coauthor', 'bookID', 'isbn', 'isbn13' as we will not consider them in building the Average Rating Prediction Model.

Changing the data type of Variables

We noticed that the datatype of the quantitative/non-categorical variables such as average average_rating, num_pages, text_reviews_count is object. The publication_date which is a date is also object.

For building a dataset suitable for predictive modelling, we changed the data type of the variables to numeric.

Converting Dates

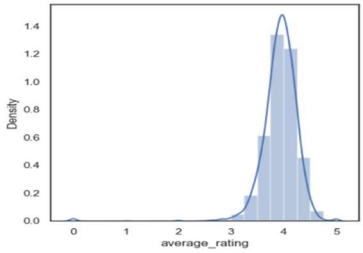
We converted the data type of publication date to datetime as the name of the variable suggests. However, during analysis, it was detected that 2 dates were unfeasible. For example: November and June had 31 days instead of 30. We corrected the respective dates and assumed that it will be the 30th day of the month under consideration.

Exploratory Data Analysis with Charts and Tables

To understand the pattern and relationship between variables, the exploratory data analysis through charts and tables will help us gain valuable insights about the dataset.

Analysis through Ratings

The distribution curve for 'Average Rating' is negatively skewed. This implies that most of the average ratings are concentrated between 3 to 5. However, as per the curve below, average rating of 5 is rare.



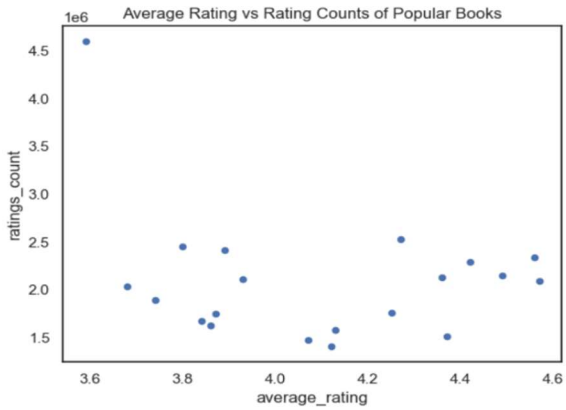
Looking at the just the top 10 books does not make sense. As there are more than 10 books with highest average rating i.e., 5.0. We pulled out a list of books with the average rating of 5.0

There were 22 books with maximum average rating (5.0).

Then we tried to find out the Top 20 most read books based on ratings count -The Popular Books among readers.

It was interesting to note that the list of popular books 20 based on ratings count is completely different from the list of books that have the highest average rating of 5.0.

We checked what is the average rating of the 20 popular books through scatter plot “Average Rating vs Ratings Count”

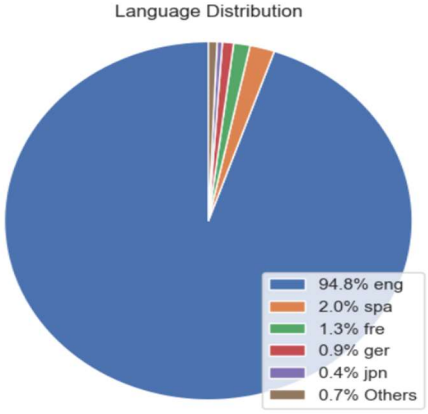


The popular books have a comparatively high rating ranging between 3.6 to 4.6

Language Distribution of Books

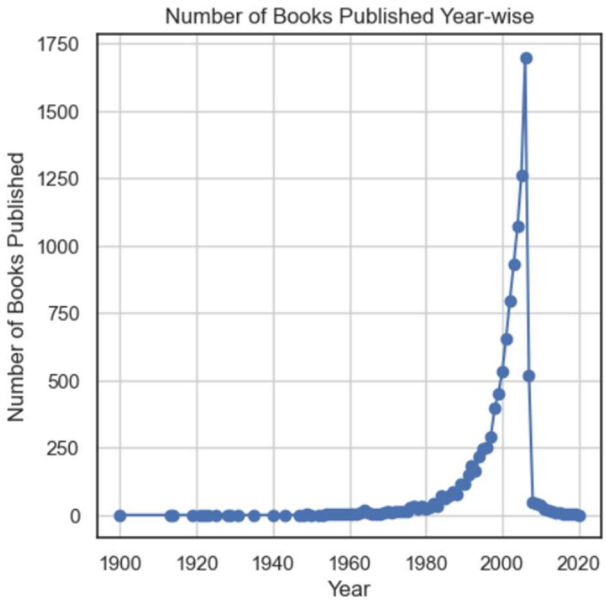
Through the pie chart provided below, we tried to understand what are the top 5 languages in which the books in the dataset are published.

Approximately, 95% of the books in the dataset are in English. The other top languages in which the books were published are Spanish(spa), French(fre), German(ger), Japanese(jpn).



Year of Publication

The hypothesis here is that with the onset of digitization and internet availability now with a major section of the population, can we conclude that less books are demanded and hence published? This may be because a certain section of the population considers watching television, films as a mean of recreation. Before digitization, when availability of television was less or could be purchased by only the elite class, people resorted to books maybe more as a mean of recreation. Through this data can we gain any such insight. Let us check:

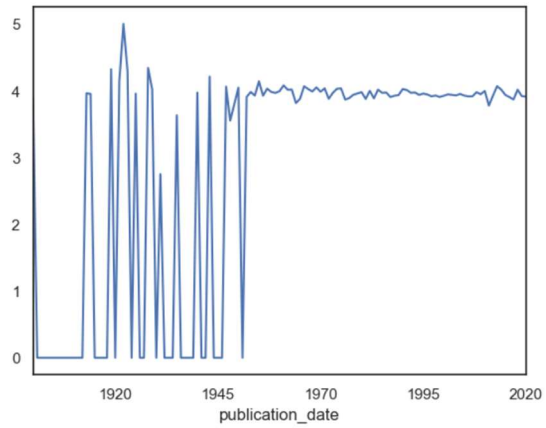


So here we see, maximum of the books present in the dataset was published between 2005 -2006.

According to our dataset, the hypothesis above is inconclusive as we do not have enough data to support it. From the above graph, we can deduce that from 1900 till 1980, the books were published at a steady rate. The very low number for books published between 1900-1960 can be attributed to loss of data for many reasons. The number of books published started increasing from ~1980. It was the highest between 2005-2006. But we suddenly saw a sharp drop after that. This can be because of a shift in demand of preferences for books to OTT platforms. A lot of leisure book readers preferred to watch a series or movie on Netflix, Amazon Prime / Hulu than read books.

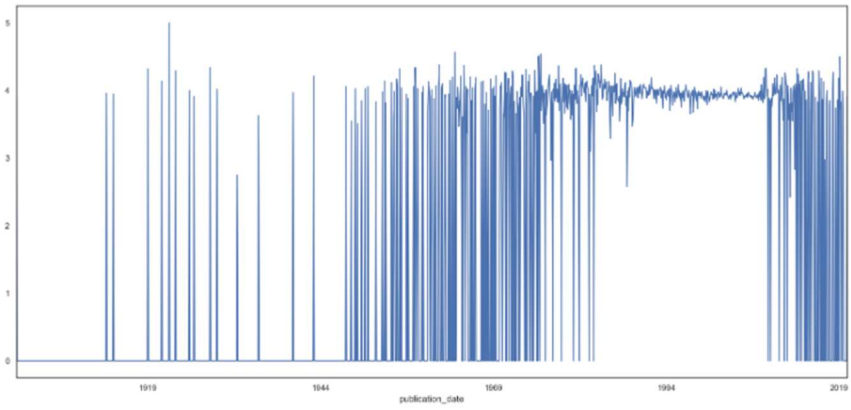
Average Rating

The average ratings of the book present in the dataset (figure below) have fluctuated a lot between 1900-1955. From 1955, the average ratings seem to be stable at 4.0 on an average. (Please note, if no books were published in a certain year, we have assumed 0 books published for that year.

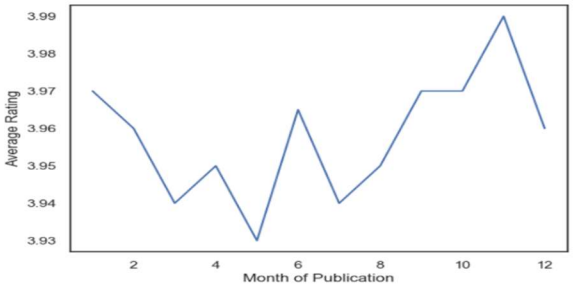


Monthly Average Rating

Through the image below can we form a hypothesis that during vacations in the month of July-August and December-January, people prefer to read more books? Can it be possible to assume that the books will have a higher rating during these four months as compared to the other months?



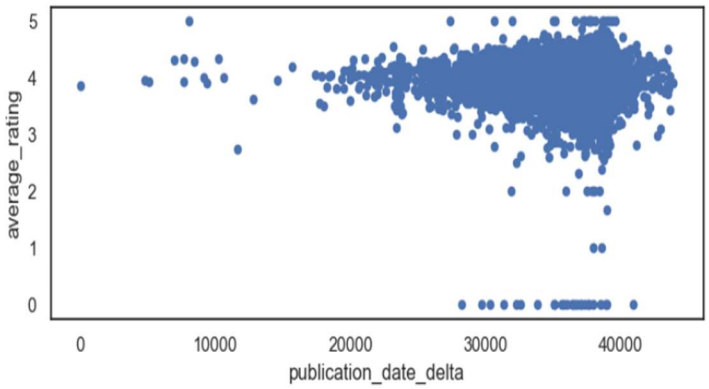
By looking at the graph below, we conclude that all the months in the year have an average rating 3.9-4.0 on an average. There is not a huge variance between the median of average rating for months. Thus, hypothesis of vacations is incorrect.



Relationship between Average Rating and other variables

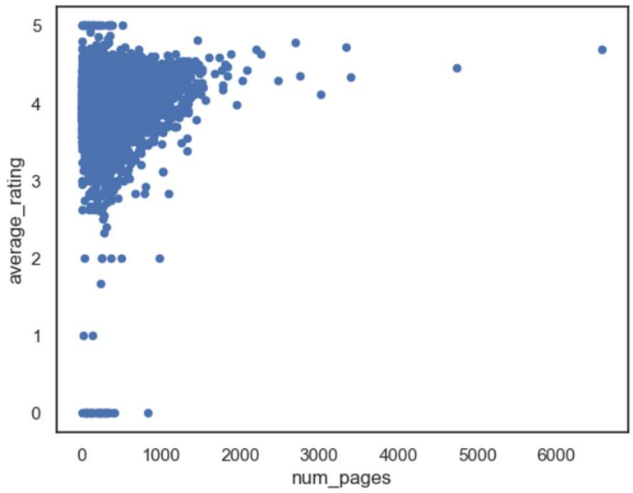
We want to analyze how average rating is related to other variables such as publication date, number of pages, number of text reviews and number of ratings counts.

Average Rating vs Publication Date



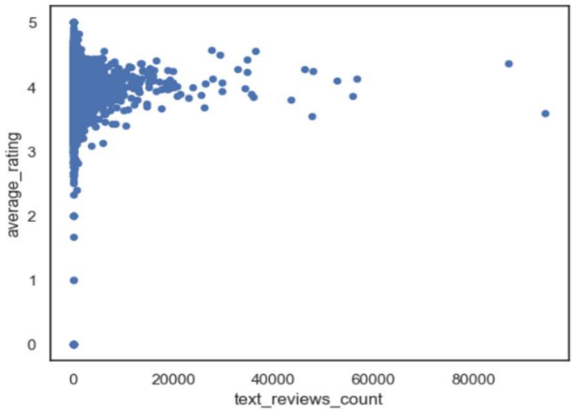
This chart is inconclusive. We cannot determine any relationship by looking at the chart. A lot observations are clustered between 30000-40000 publication date delta.

Average Rating vs Number of Pages



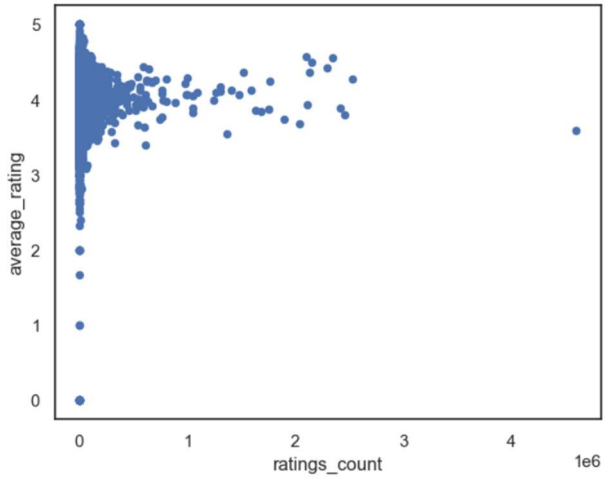
Through this chart, we can still conclude that there exists a weak positive relationship between number of pages and average rating. The higher the number of pages, average ratings are also increasing.

Average Rating vs Text Review Count



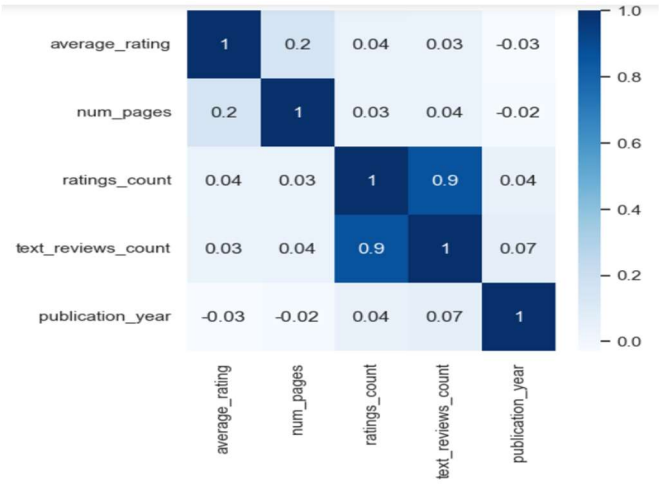
As per the above chart, there exists a weak positive relationship between ratings count and average rating. This sets the stage for looking at correlation between the variables.

Average Rating VS Ratings Count



As per the above chart, there exists a weak positive relationship between ratings count and average rating. This sets the stage for looking at correlation between the variables.

Correlation Table



From the correlation table above, we deduced that text reviews count has a very high positive relationship with ratings count. Average rating has weak positive relationship with number of pages, text revies count and ratings count. It also has a weak negative relationship with publication year.

Filtering Data and Feature Engineering

1. Separate out all the co-authors. Co-authors are separated by a /.

2. We take the highest rated records and filter out records with rating > 4.

3. Get the top 50 authors and co-authors from them

Here, we have derived the list of the top 50 authors. Firstly, we filtered the data frame based on ratings count in descending order and then filtered the books data having average rating greater than 4. Finally, from this filtered dataset we generated the list of Top 50 authors.

The double filtering based on ratings count and average rating denotes the authors selected in the list must have written books with substantial ratings count and average rating greater than 4.0.

Later in this notebook, we will use this list of top 50 authors to develop a feature 'Is Top Author' for building the machine learning model.

4. similarly calculate the top 50 publishers

We have used a similar approach that was used to generate the list of top authors here to generate the list of top publishers.

Later in this notebook, we will use this list of top 50 publishers to develop a feature 'Is Top Publisher' for building the machine learning model.

5. Construct two additional columns that is a part of the feature engineering

The columns are based on the following understand:

- Whether given author belongs to a group of top 50 rated authors
- Whether given publisher belongs to a group of top 50 rated publisher

We further clean the dataset by removing the unwanted columns.

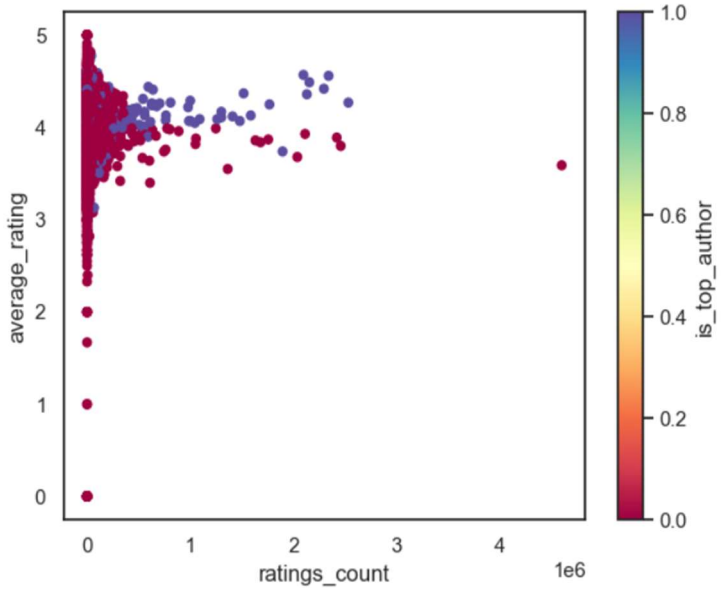
6. In the above step, we drop columns the author's name, publisher name and title from the original dataset

7. We created 2 additional columns: is_top_author, is_top_publisher. This is a part of our feature engineering process.

8. We created a composite feature named 'Top Author Publisher' which denotes the datapoint belongs to a top author and publisher.

Data Analysis and relationship analysis through graph and charts after filtering and featuring

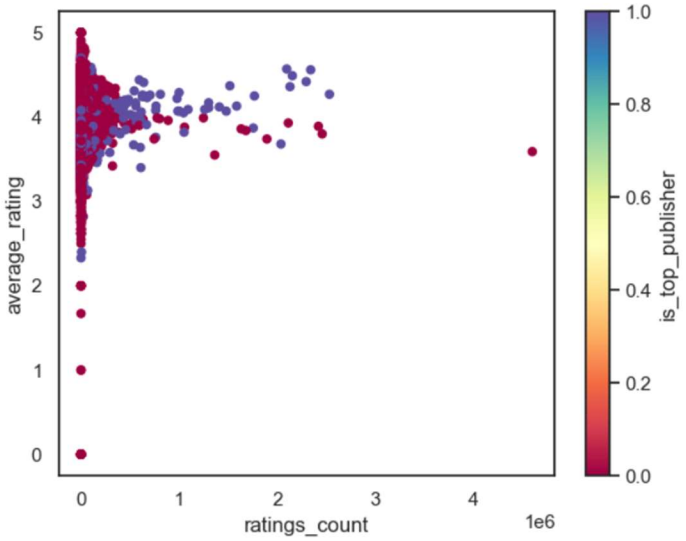
Average Rating VS Rating Count (Top Author)



The bubbles in the above chart represent if the book belongs to a top author or not. The purple bubbles represent data points of top author and the red represent data points of not a top author.

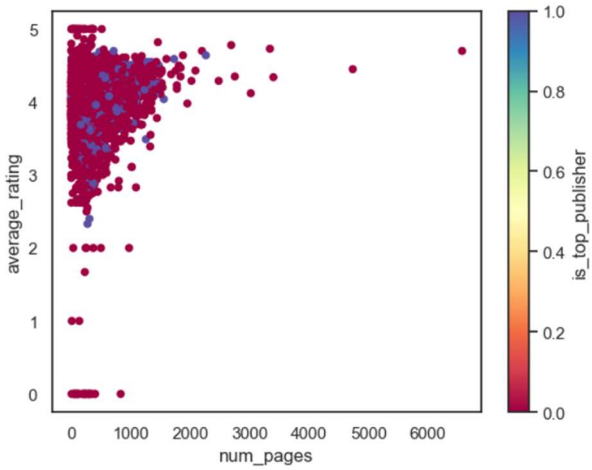
We can see that the average rating and ratings count have a positive relationship for top authors.

Average Rating VS Rating Count (Top Publisher)



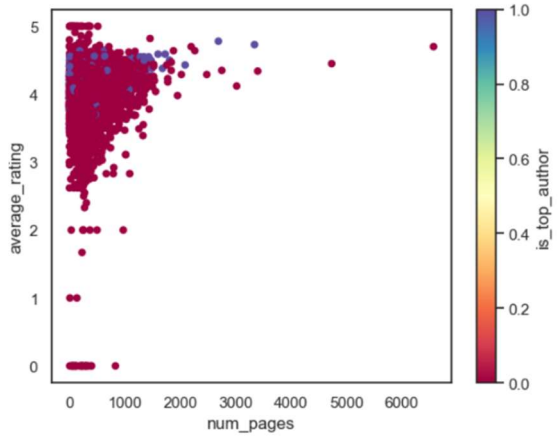
For only top publishers, there exists a positive relationship between average rating and ratings count. This means higher the ratings count, higher is the average rating.

Average Rating VS Number of Pages (Top Publisher)



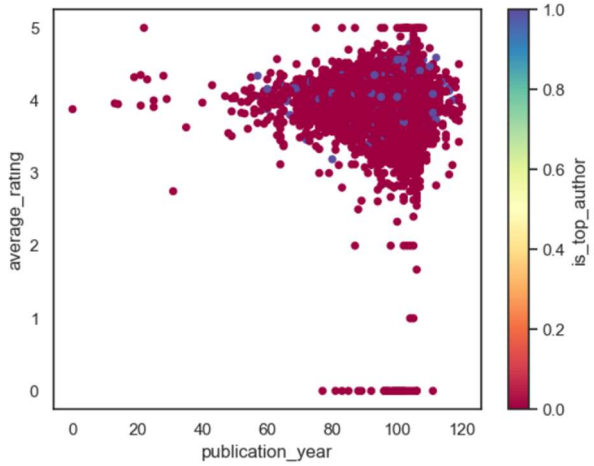
The above graph also denotes for the top publishers, higher the number of pages, higher is the average rating.

Average Rating VS Number of Pages (Top Author)



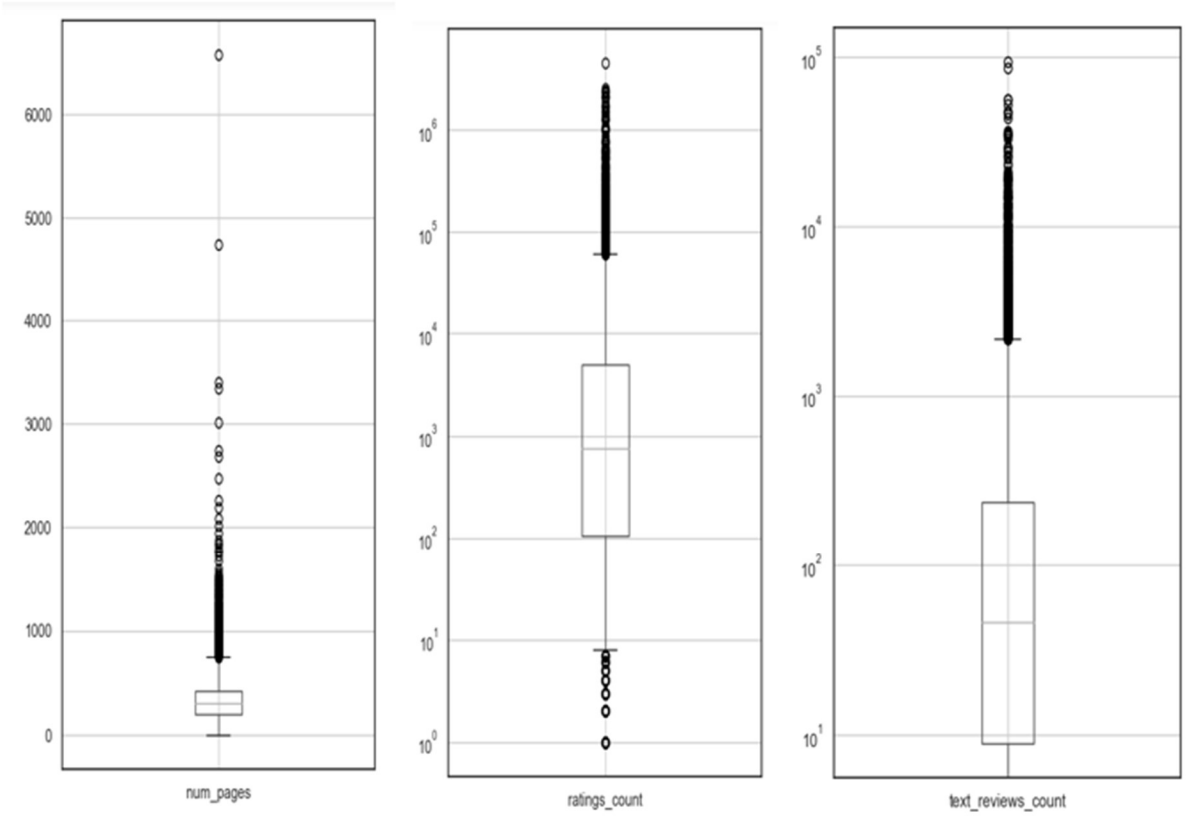
The above chart is inconclusive. We cannot determine any relationship between average rating and number of pages.

Average Rating VS Publication Year (Top Author)



The above chart is inconclusive. We cannot determine any relationship between average rating and publication year.

Removing Outliers



The 'eligible_df' was formed by removing extreme values or outliers from number of pages, ratings count and text reviews count.

The removal of outliers is an important step in model building. The presence of outliers will affect the accuracy of the model.

For example: While going through the dataset (check the dataframe 'check_outliers'), it was noticed that the ratings count for some books were only 1. This is a biased data. We cannot feed this data to our machine learning model during the supervised learning process. This biased data will result in data anomaly or wrong prediction.

We have also dropped the 'title', 'language code' and 'year' column from the eligible df.

Model Building

We created two data frames X and Y. X represents the independent variables. Y is basically average rating which we want to predict with the help of our model. Hence, Y data frame consists of the dependent variable

We did not take any scaler as it did not have any effect on the accuracy of the model.

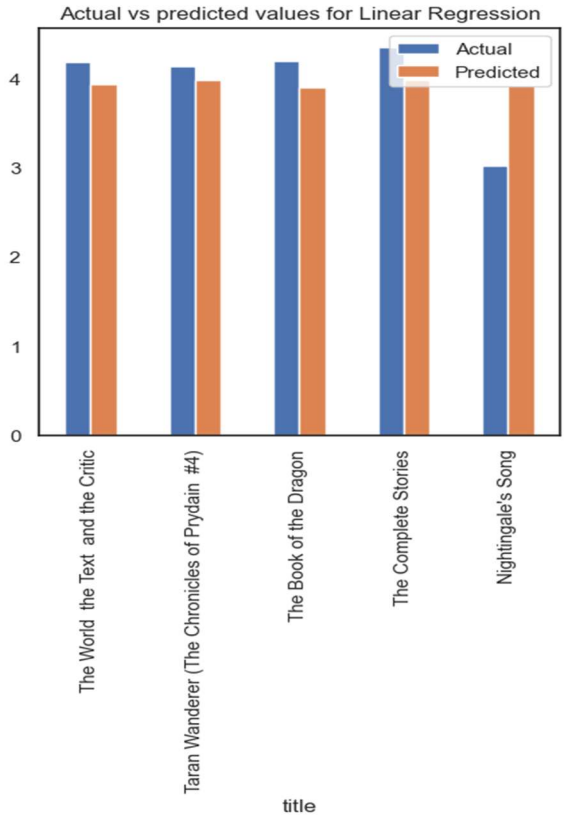
Linear regression Baseline

To get quick results and to see that our feature engineering worked. We immediately fed this to a linear regression model. This can act as a quick baseline.

We see the Mean Squared Error (MSE) is 8%. The more accurate a model, MSE tends to 0. Our objective is to minimize MSE.

```
MAE: 0.21
MSE: 0.08
RMSE: 0.28
```

In the graph provided below, we provide a visual representation of the y-value predicted from the linear regression model and the actual y-value present in the data set for 5 samples.



KFold Cross validation Method (k=4)

We use KFold cross validation method which gives us avg accuracy = -0.08 or m.s.error of 8%.

Please note, in scikit scores are always supposed to be maximized. Hence, we use negative mean square error instead of mean square error. We need to maximize this error meaning, closer to 0 better the model.

Comparative analysis of ML Algorithms

Now we compared the Linear Regression Model against XG Boost, Random Forest, and Decision Tree.

We have compared the average loss across different machine learning models mentioned above to get a comprehensive outlook and optimize our results.

For fair comparison we gave the same split of X_train and Y_train to all the models.

The loss of Random Forest Model and Linear Regression Model is the lowest as compared to the other models

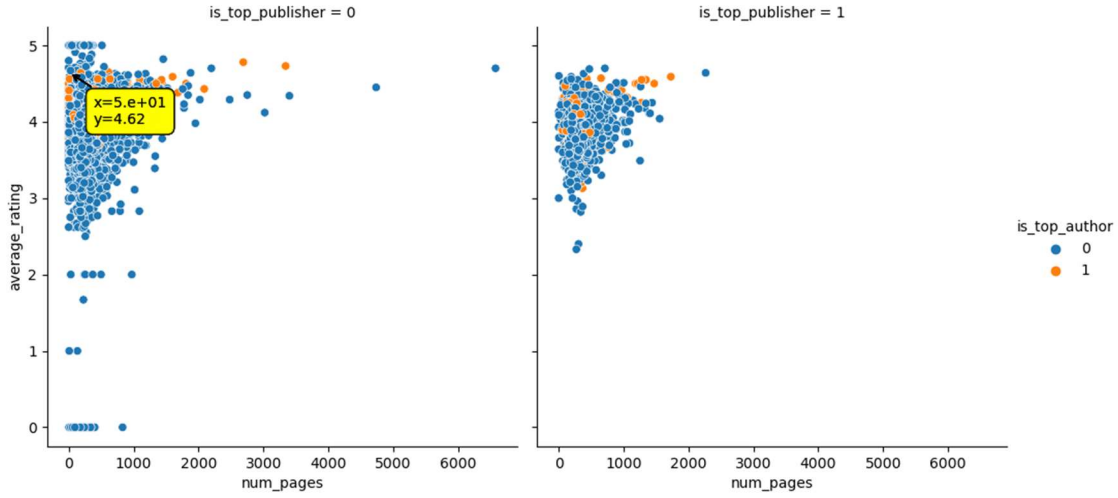
As compared to all the models, linear regression model and random forest model have higher accuracy i.e., the average loss for both the models is 0.08.

```
Fold = 1
The average loss of linear regression is 0.08
The average loss of Decsion Tree is 0.15
The average loss of XG Boost is 0.09000000357627869
The average loss of Random Forest is 0.07
Fold = 2
The average loss of linear regression is 0.09
The average loss of Decsion Tree is 0.16
The average loss of XG Boost is 0.10000000149011612
The average loss of Random Forest is 0.08
Fold = 3
The average loss of linear regression is 0.08
The average loss of Decsion Tree is 0.15
The average loss of XG Boost is 0.09000000357627869
The average loss of Random Forest is 0.08
Fold = 4
The average loss of linear regression is 0.08
The average loss of Decsion Tree is 0.16
The average loss of XG Boost is 0.10000000149011612
The average loss of Random Forest is 0.08

----------------------Consolidated Results-----------------------------
The average loss of linear regression is 0.08
The average loss of XG Boost is 0.09000000357627869
```

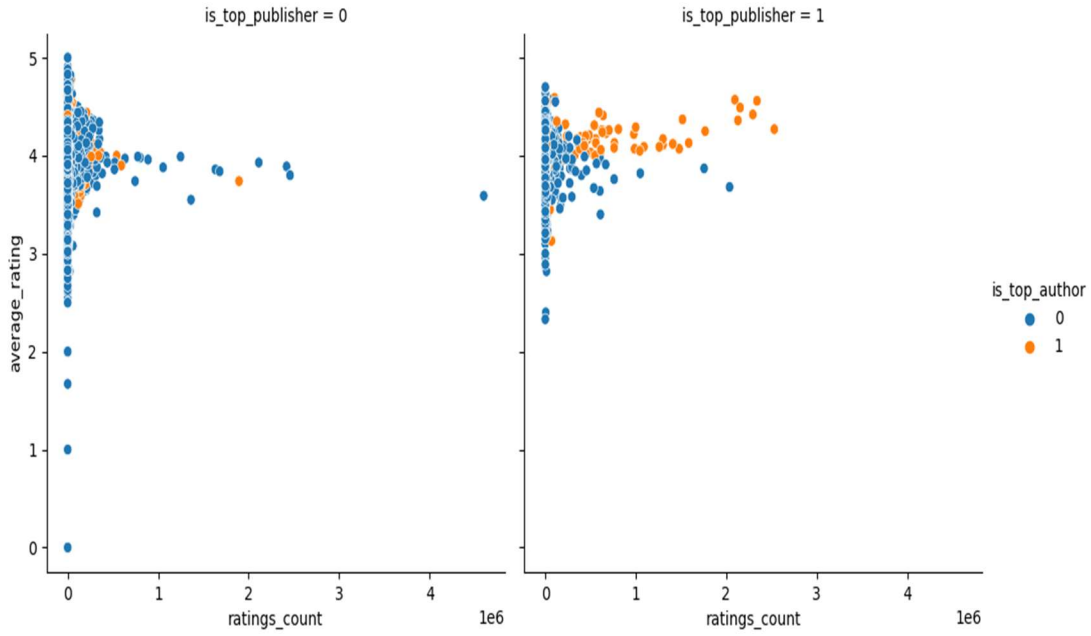
```
The average loss of Random Forest is 0.08
The average loss of Decision Tree is 0.16

----Some Predicted values for Linear Regression ----
                    title                                    Actual  Predicted
2   Harry Potter and the Chamber of Secrets (Harry...    4.42   4.046871
5   Unauthorized Harry Potter Book Seven News: "Ha...    3.74   3.893591
7   The Ultimate Hitchhiker's Guide: Five Complete...    4.38   4.014864
9   The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22   3.886748
10  The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22   3.843145
----Some Predicted values for Decision tree ----
                    title                                    Actual  Predicted
2   Harry Potter and the Chamber of Secrets (Harry...    4.42       4.10
5   Unauthorized Harry Potter Book Seven News: "Ha...    3.74       3.85
7   The Ultimate Hitchhiker's Guide: Five Complete...    4.38       3.92
9   The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22       3.75
10  The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22       3.91
----Some Predicted values for Random Forest ----
                    title                                    Actual  Predicted
2   Harry Potter and the Chamber of Secrets (Harry...    4.42   3.945866
5   Unauthorized Harry Potter Book Seven News: "Ha...    3.74   3.886244
7   The Ultimate Hitchhiker's Guide: Five Complete...    4.38   4.082098
9   The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22   3.839041
10  The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22   4.005677
----Some Predicted values for XG Boost ----
                    title                                    Actual  Predicted
2   Harry Potter and the Chamber of Secrets (Harry...    4.42   3.970488
5   Unauthorized Harry Potter Book Seven News: "Ha...    3.74   3.908060
7   The Ultimate Hitchhiker's Guide: Five Complete...    4.38   4.079884
9   The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22   3.760745
10  The Hitchhiker's Guide to the Galaxy (Hitchhik...    4.22   3.880234
```

Interactive Graphs and Charts Showing Relationship between the Features

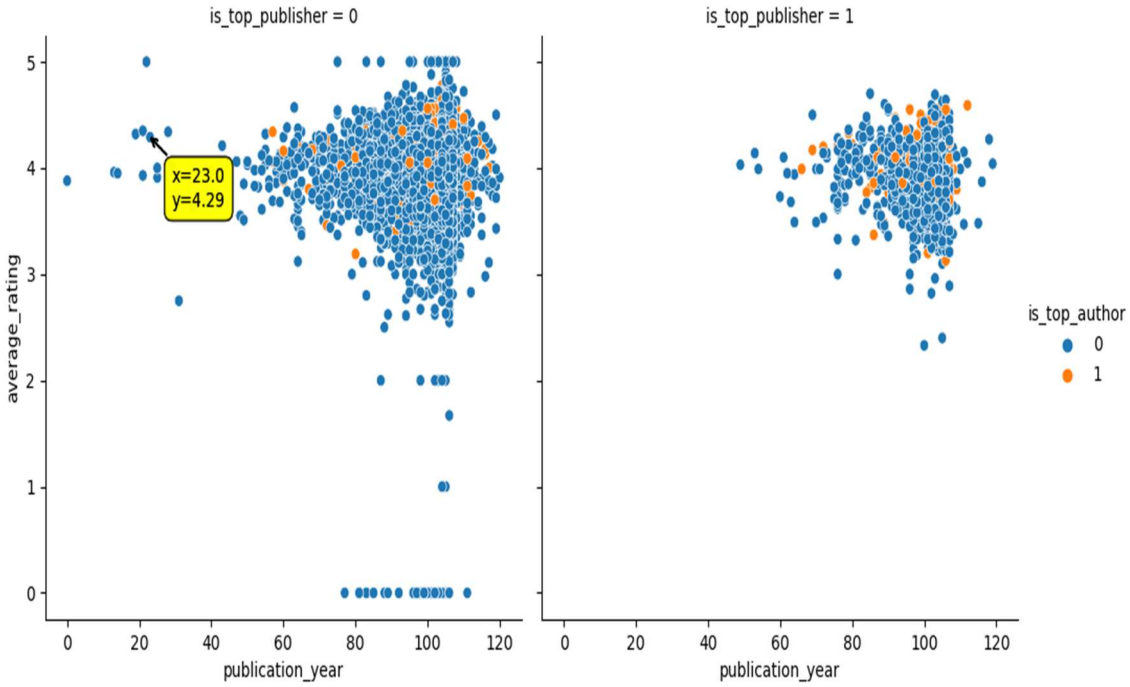
Average Rating vs Num of Pages (top publisher and top author)



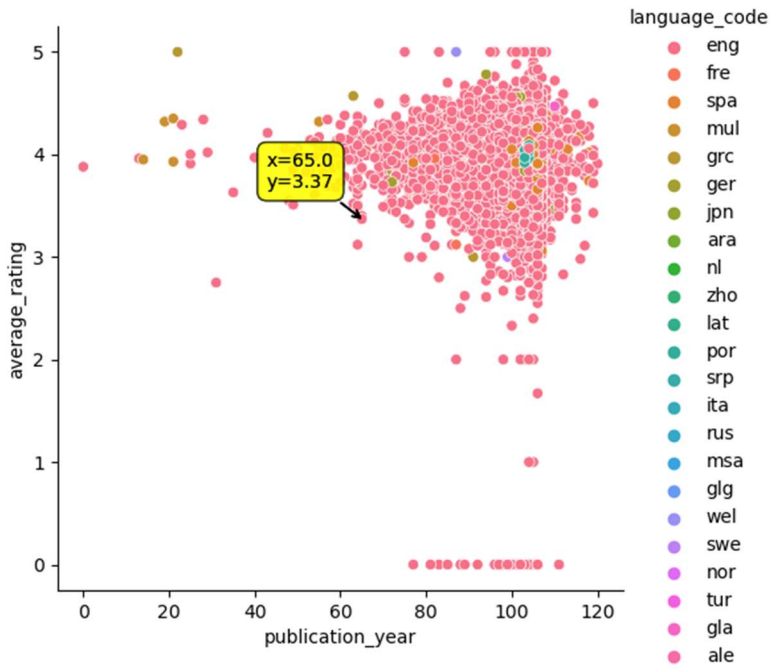
Average Rating vs Ratings Count (top publisher and top author)



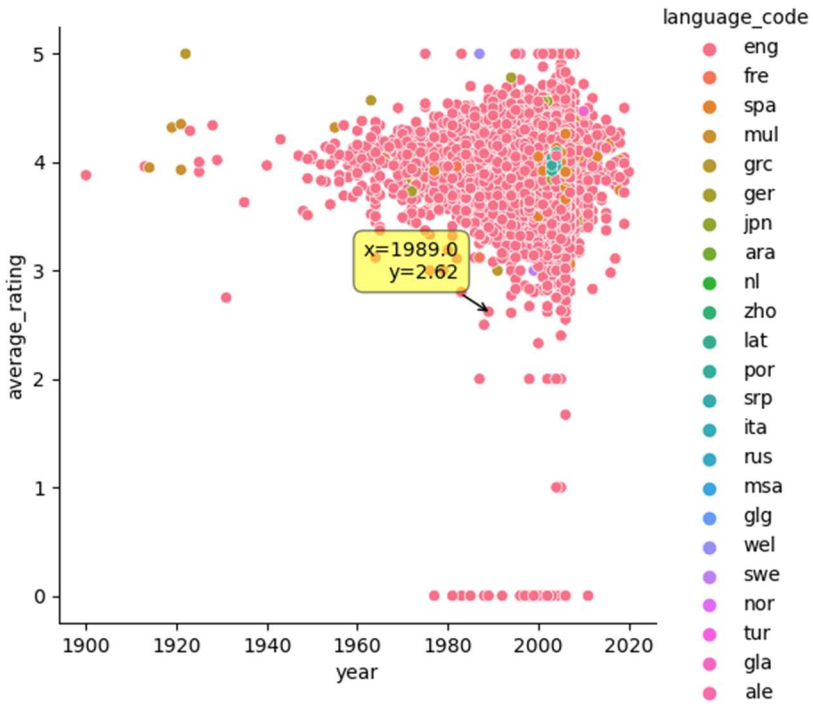
Average Rating vs Publication Year (top publisher and top author)



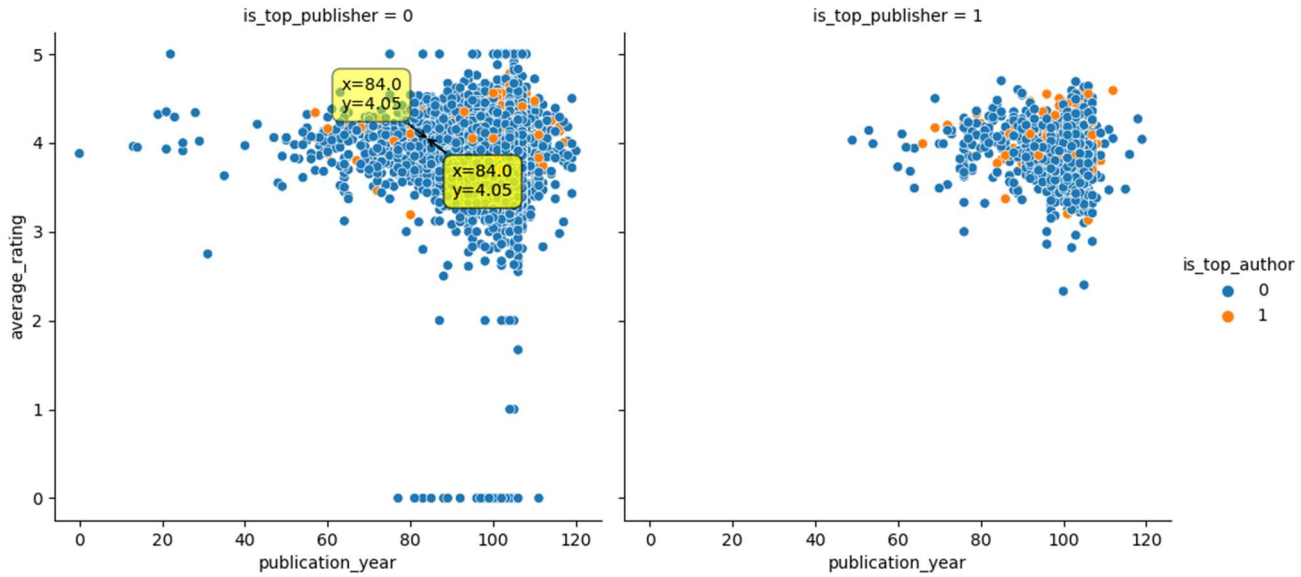
Average Rating vs Publication Year (language code)



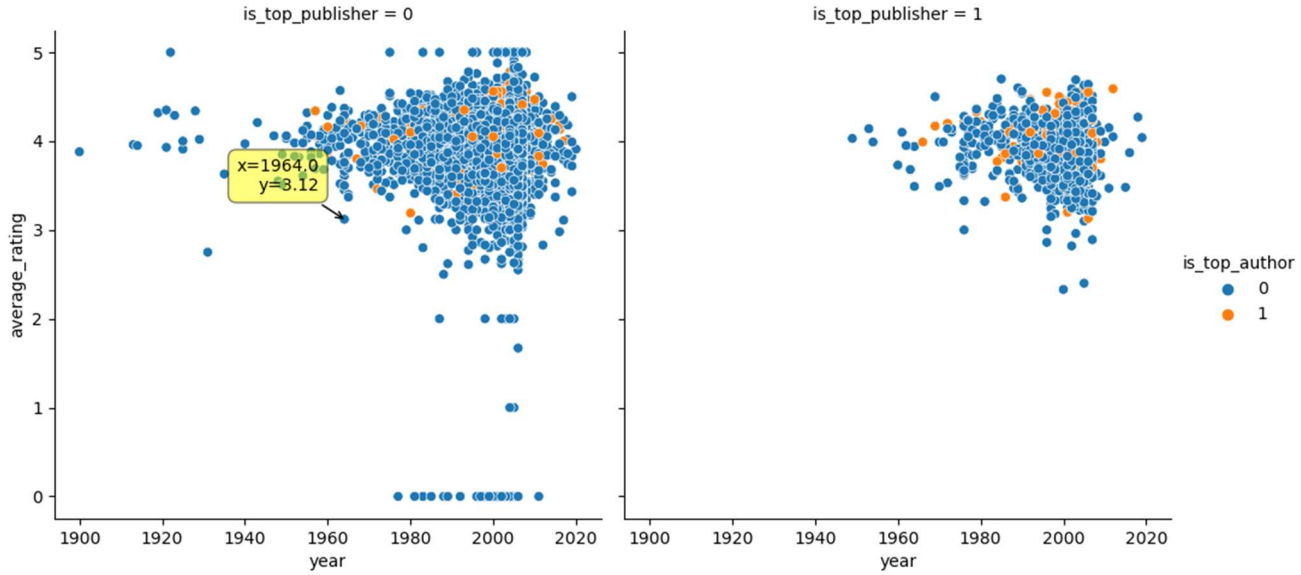
Average Rating vs Year (language code)



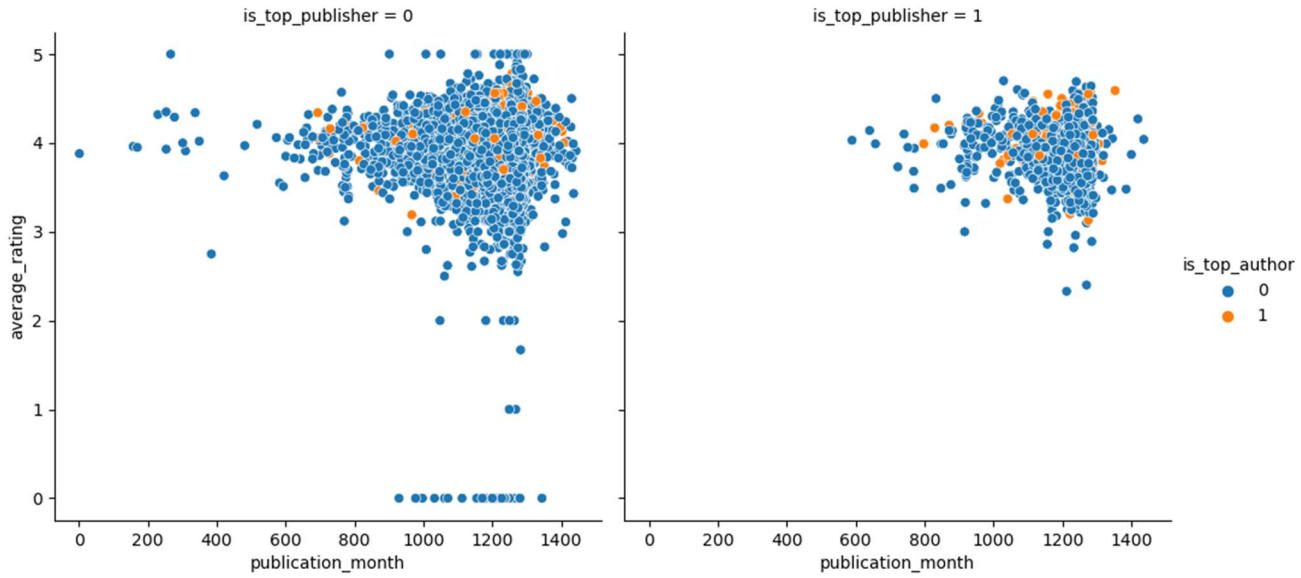
Average Rating vs Publication year (top publisher and top author)



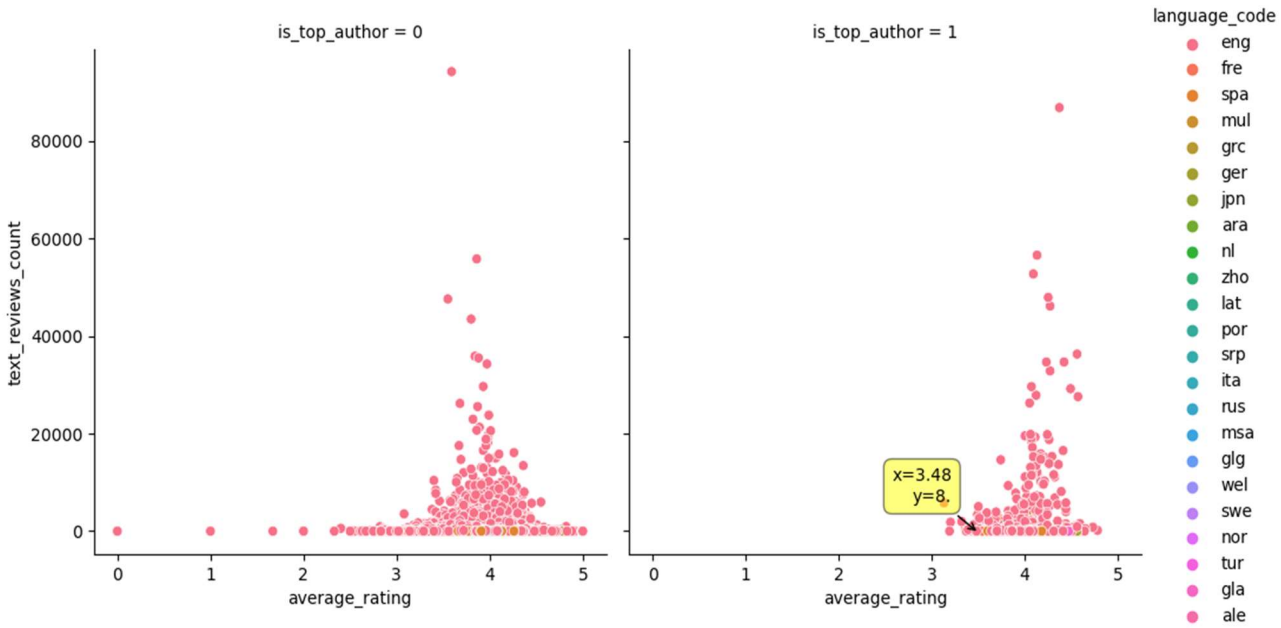
Average Rating vs Year (top publisher and top author)



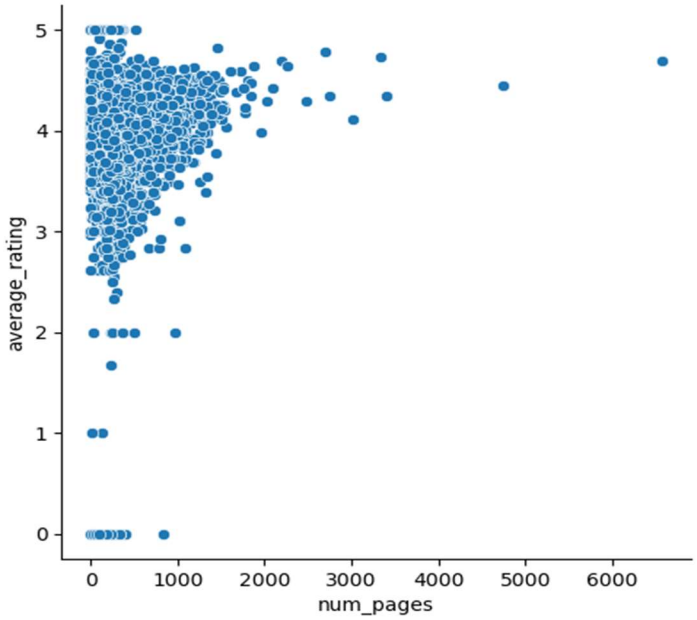
Average Rating vs Publication month (top publisher and top author)



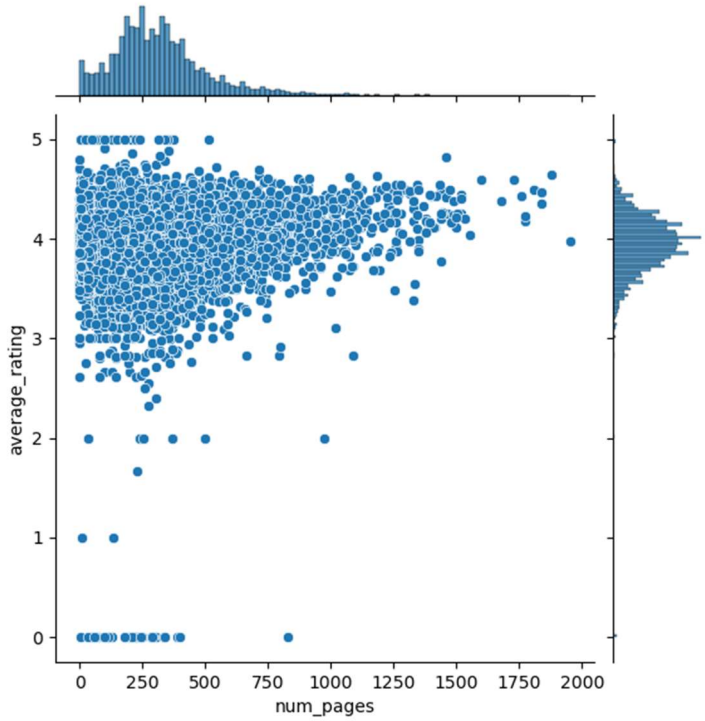
Text review count vs Average rating (top author, language code)



Relationship between High Rating and Number of Pages



Most of the books with number of pages<2000 have an average rating between 3-4. This graph doesn't help in concluding anything. Let's explore a bit more for number of pages<2000



Based on the joint plot above, we can infer that readers prefer reading books between 200-300 pages and rate them between 3-5

Conclusion

Our model predicts, average rating of given books in the dataset, for which Random Forest and Linear Regression models performed the best. Both got Mean Squared Error of 0.08.

The features we considered for modelling:

1. `num_pages`
2. `ratings_count`
3. `text_reviews_count`
4. `publication_year`
5. `publication_month`
6. `is_top_author`
7. `is_top_publisher`
8. `top_author_publisher`