Capstone Project-IV

Book Recommendation System

Based
on
Unsupervised Machine Learning

Presented
by

SUDHANSHU KUMAR

Contents

Problem statement

- The goal of recommendation systems is to generate a list of suggestions for a user.
- During the last few years, with the rise in digital revolution, people switching towards OTT platform like YouTube, Amazon, Netflix, and many other such web services, recommender systems have been more and more important in our lives because it provides more customized and relevant content.
- The main objective is to build a recommendation system that would suggest appropriate books to users based on their interests and book popularity.

Data Summary

Books.csv, Ratings.csv, User.csv

Three datasets

Users dataset

```
users_df.shape
```

```
(278858, 3)
```

1. User-ID
2. Location (city, state, country)
3. Age

Rating dataset

```
ratings_df.shape
```

```
(1149780, 3)
```

1. User-ID
2. ISBN
3. Book-Rating

Book dataset

```
books_df.shape
```

```
(271360, 8)
```

1. ISBN
2. Book-Title
3. Book author
4. Year-Of-Publication
5. Publisher
6. Image-URL-S
7. Image-URL-M
8. Image-URL-L

Checking Null Data and Duplicates

```
data.isnull().sum().sort_values(ascending=False)
```

```
data[data.duplicated()]
```

No Duplicates

Users dataset

Data=users_df

Age	110762
User-ID	0
Location	0

- Checking outlier
  - ✓ We checked maximum and minimum age in datasets.
  - ✓ Replacing age all age greater than 100 with NaN (outlier treatment).
- Replacing NaN value with mean of all the age .

Rating dataset

Data=ratings_df

User-ID	0
ISBN	0
rating	0

- There is no missing values

No Duplicates

No Duplicates

Book dataset

Data=books_df

Image-URL-L	3
Publisher	2
author	1
ISBN	0
title	0
publication_year	0
Image-URL-S	0
Image-URL-M	0

- Fetching book information through book URL we filled the missing value for author and Publishers features.
- Drop Image-URL-L columns

Outlier Treatment

Users dataset

Data=users_df

For Age columns:

- We checked maximum and minimum age in datasets.
- Replacing age all age greater than 100 with NaN (outlier treatment).
- Replacing NaN value with mean of all the age .

Rating dataset

Data=ratings_df

- There is no outlier.

Book dataset

Data=books_df

For publication year columns:

- minimum year is 0 and maximum year is 2050.
- We calculate mode of publication year columns.
- Replacing publication year 0 and publication year after 2021 with mode.

Feature Engineering

Users dataset

Location (city, state, country)

```
users_df['country']=users_df['Location'].str.split(',').str.get(-1)
```

Country

Users datasets
Books datasets
Ratings datasets

Combine three datasets using dataframe.merge

Combined dataset (total_df)

total_df

Age

```
bins= [3,12,20,50,100]
labels = ['Kid','Teen','Adult','old']
total_df['AgeGroup'] = pd.cut(total_df['Age'], bins=bins, labels=labels, right=False)
```

AgeGroup
Kid(3-12)
Teen(13-20)
Adult(21-50)
Old(51-100)

total_df

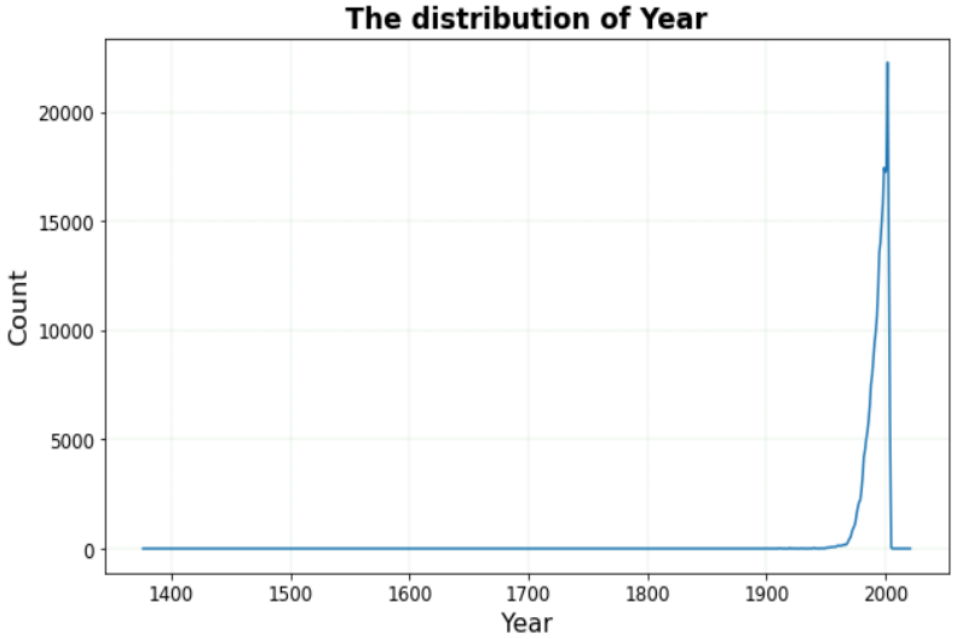
Rating

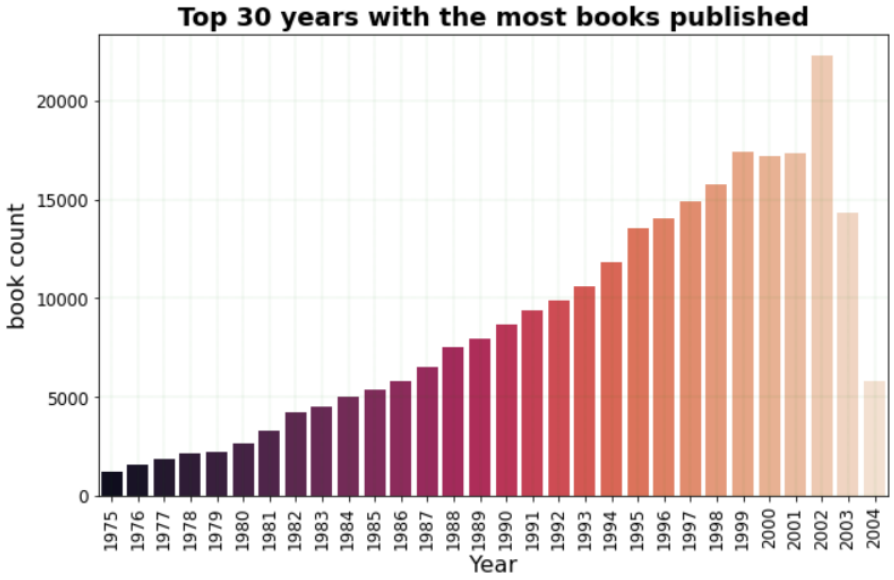
```
bins= [0,1,10]
labels = ['Implicit','Explicit']
total_df['rating_cat'] = pd.cut(total_df['rating'], bins=bins, labels=labels, right=False)
```

Rating_category
0- Implicit
1- 10-Explicit

Exploratory data analysis

Publication year and Book count



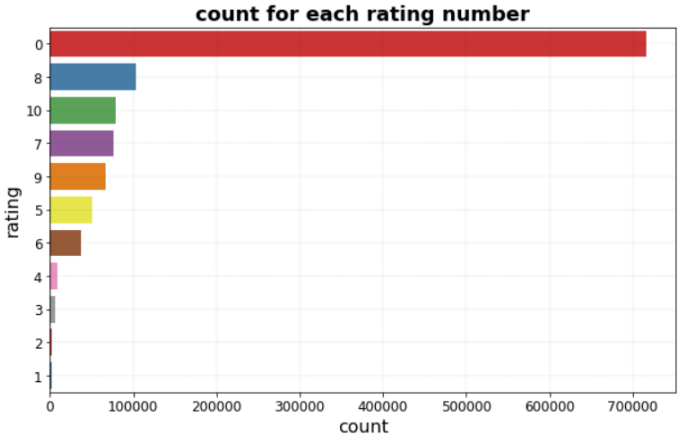


Most of the books are published between 1950 to 2002

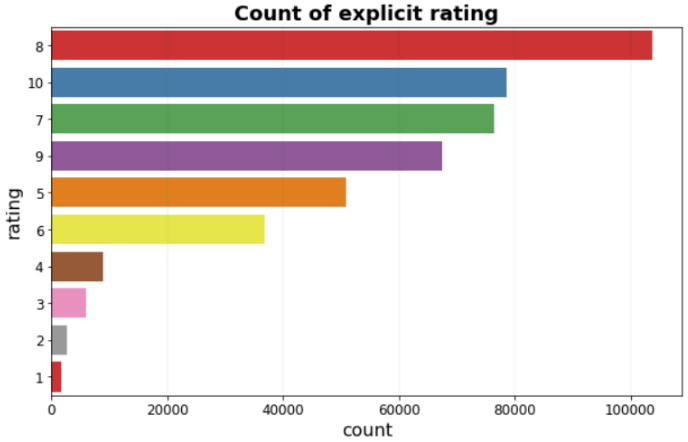
In this diagram, it is found that book count increases gradually from 1975 to 2002, and reaches to its maximum value in 2002 and then decreases.

Exploratory data analysis

Rating vs Book count

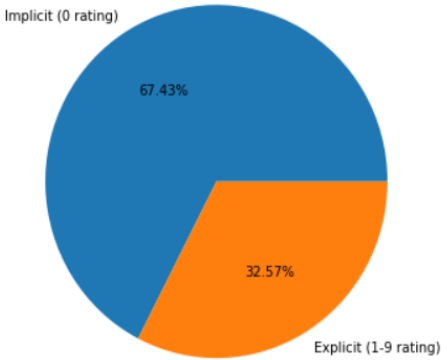


- book count for 0 rating is large as compare to others ratings,



When we exclude 0 rating:

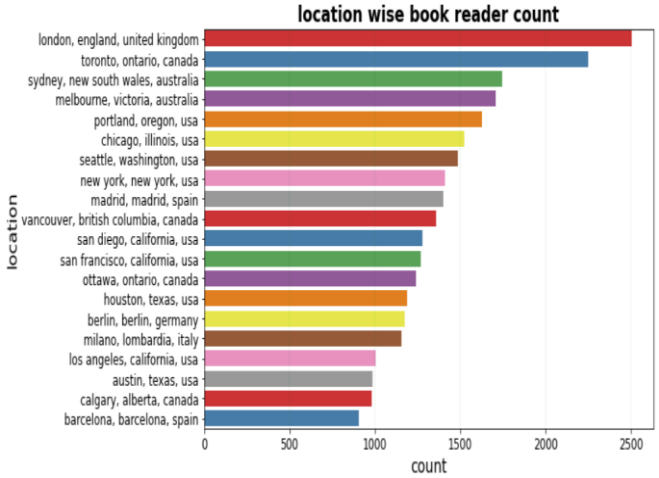
- most of the user rated 8 followed by 10 and 7.
- very less number of user rated 1.



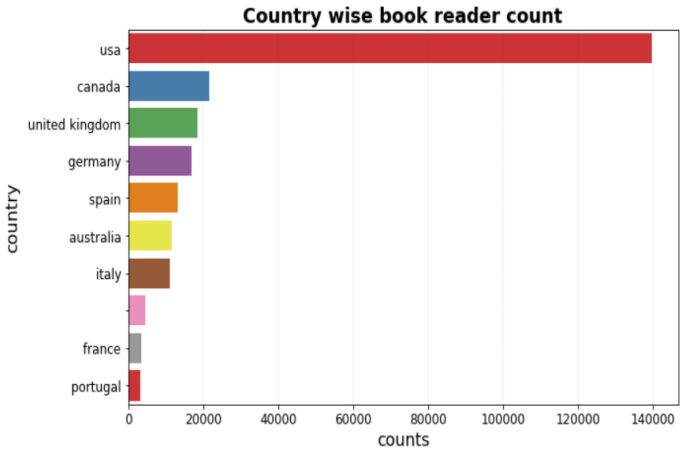
- 67.43% users give 0 rating for books and only 32.57 % user are positively rated the books.

Exploratory data analysis

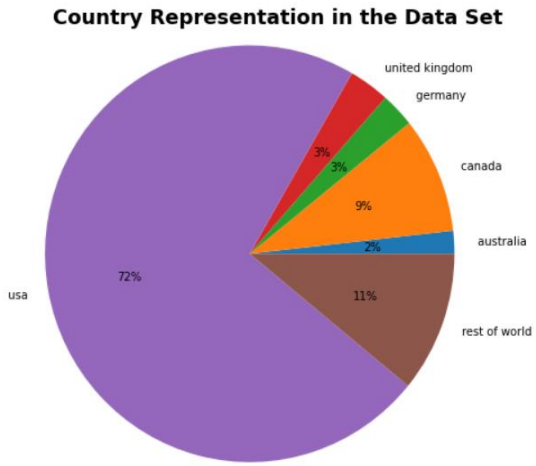
Location vs User count



- London, England, united kingdom is the most popular place for book reader followed by Canada.
- out of top 20 location, 9 location from USA followed by Canada.

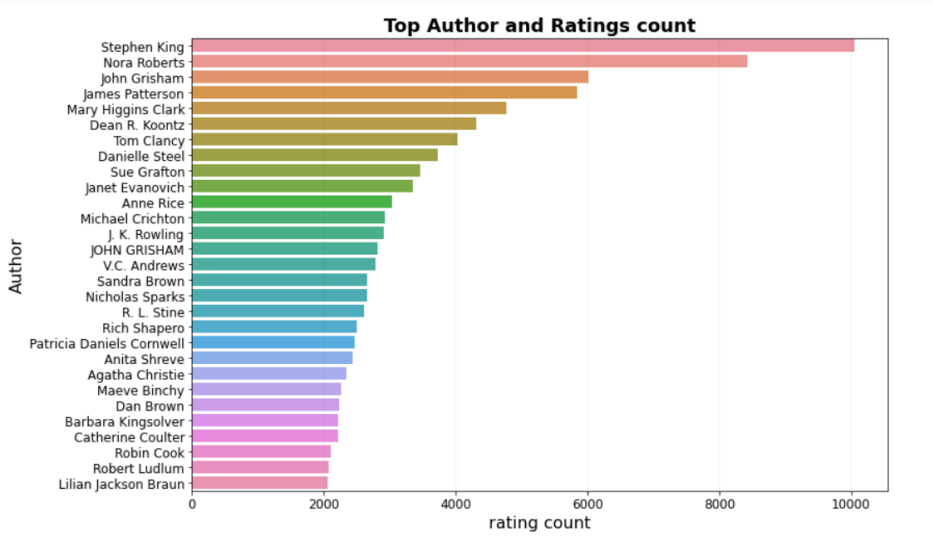


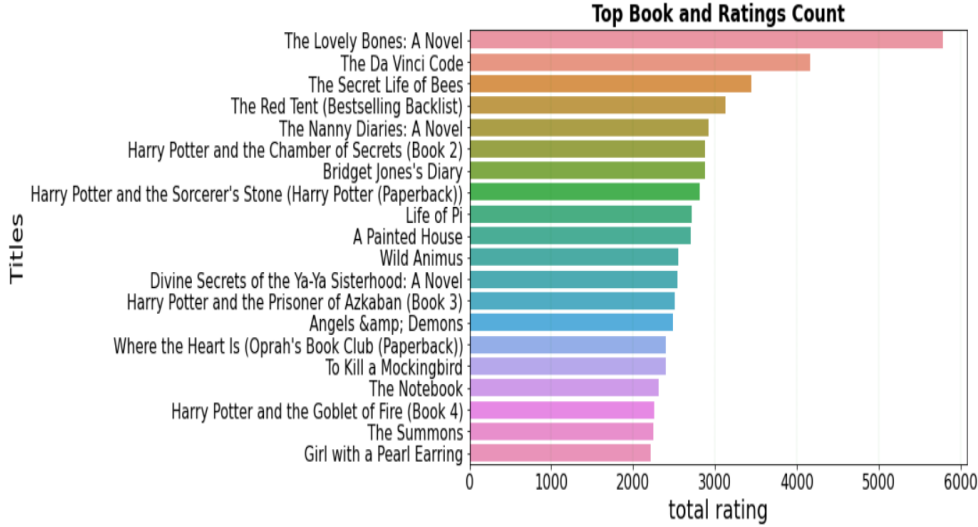
- When we extracted country name from location columns, it is found that most of the book reader from USA followed by Canada and united kingdom.



72% of books reader from usa and 9% from canada.

Exploratory data analysis

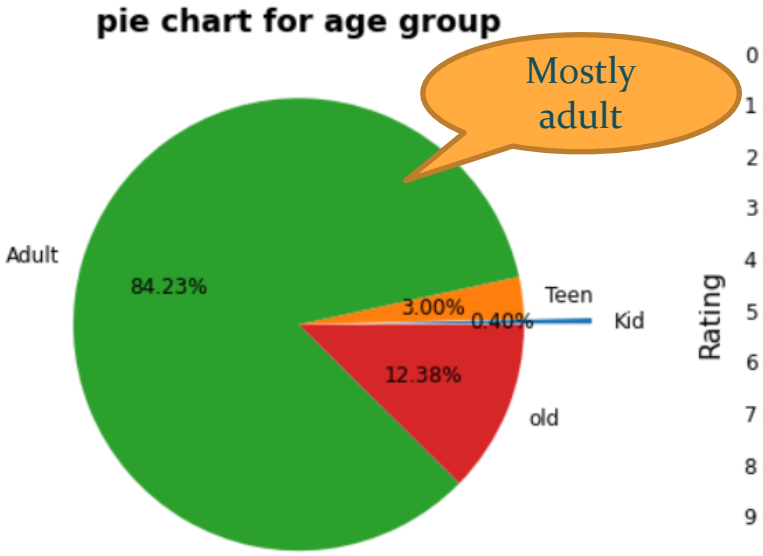




- when we consider only those author whom book rating count more than 2000 then Stephans King is most popular author followed by Nora Roberts.

- The lovely Bones: A Novel is most rated books followed by The Da Vinci Code.

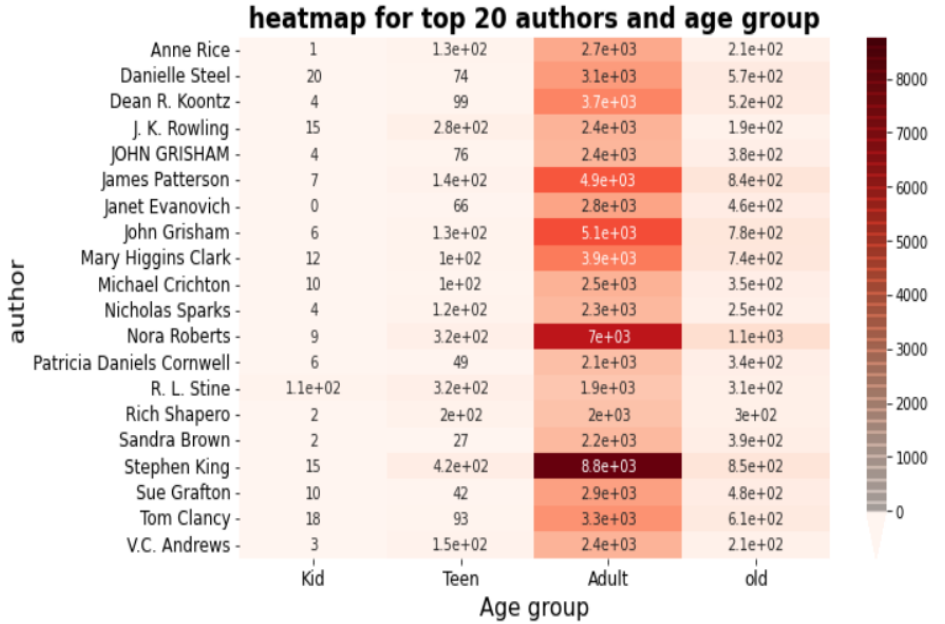
Exploratory data analysis

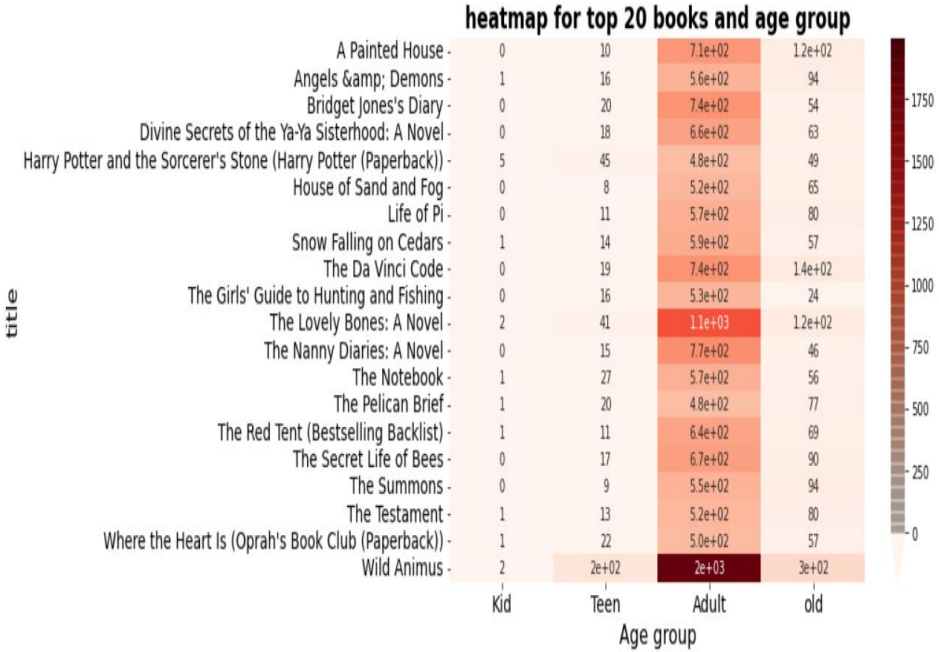


Adult=84.23%
Old=12.38%
Teen =3%
Kid=0.4%

- Most of the user who rate the book are from adult age group.
- the age group who rate down are from adult age group.

Exploratory data analysis





- Most of the book reader from adult age group.
- Stephan king rated by most of the users.

- wild animus is popular among all the four age group.

Model Building

1. **Weighted average using IMDB formula**

$$Weighted\ \ Rating = \frac{v}{v+m} \times R + \frac{m}{v+m} \times C$$

Where,

v is the number of votes for the books;
m is the minimum votes required to be listed in the chart;
R is the average rating of the book; and
C is the mean vote across the whole report.

Assumption:

We are considering explicit rating (rating between 1 to 10)

Step involved:

- Calculate the value of m and c
- Create a new feature (score) and apply weighted rating formula.
- Recommended book are those book which have highest score value.

Model Building

	title	author	AgeGroup	No_Of_Users_Rated	Rating_average	Score
	Harry Potter and the Chamber of Secrets (Book 2)	J. K. Rowling	old	118	8.050847	7.755712
	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	J. K. Rowling	Adult	161	7.937888	7.729911
	The Secret Life of Bees	Sue Monk Kidd	Adult	221	7.850679	7.705022
	The Bean Trees	Barbara Kingsolver	Adult	88	8.034091	7.681936
	Fahrenheit 451	RAY BRADBURY	Adult	109	7.935780	7.661311
	The Joy Luck Club	Amy Tan	Adult	162	7.839506	7.656407
	To Kill a Mockingbird	Harper Lee	Adult	109	7.926606	7.655099
	The Lovely Bones: A Novel	Alice Sebold	Teen	559	7.704830	7.652162
	Ender's Game (Ender Wiggins Saga (Paperback))	Orson Scott Card	Adult	70	8.057143	7.643205
	Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. Rowling	Adult	64	8.093750	7.641991

- out of top 10 books, only 1 books are popular among teen rest of them are popular for adult and old age group.
- Harry Potter and the Chamber of Secrets (Book 2) is getting maximum score.
- out of top 10 books,3 books are written by J. K. Rowling is more popular.
- The Lovely Bones: A Novel written by Alice sebold is more popular among teen and it is rated by maximum number of users.

Model Building

2. Recommendation based on rating count

In this case we are considering those book which have rated by minimum 30 users and average rating more than 5.

top 5 book according to user rating

	User-ID	title	rating count	avg_rating
0	70396	Free	56	8.018
1	233398	Where the Sidewalk Ends : Poems and Drawings	33	7.121
2	112160	Love You Forever	48	6.792
3	64207	Falling Up	39	6.744
4	96823	Johnny Got His Gun	34	6.588

Model Building

3. Model based collaborative filtering

- We are considering 10% of frequently explicitly rated books.
- Performing 5 fold cross validation to check the performance of model
- Optimal model are those which are more accurate and lower training/testing time.

Model Performance:

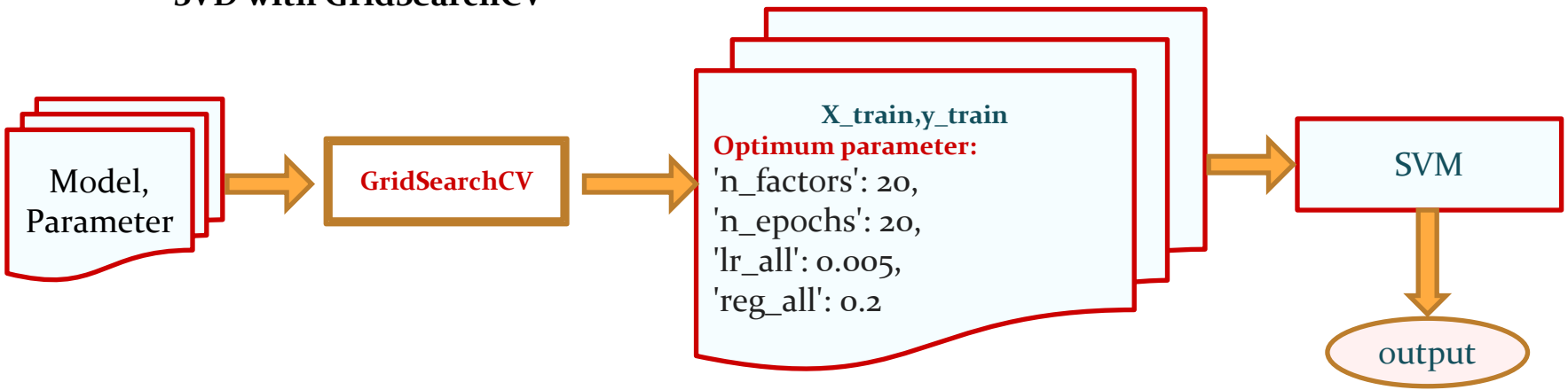
	MAE	RMSE	fit_time	test_time
matrix_factorization.SVD	1.133	1.462	8.127	0.344
matrix_factorization.NMF	1.835	2.197	12.289	0.298

From above table it is observe that:

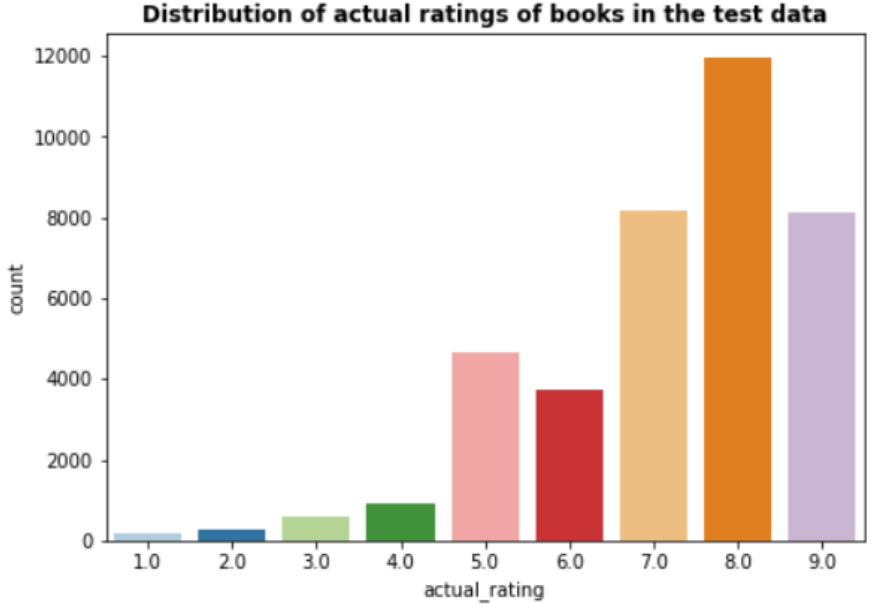
- For the given dataset much better results can be obtained with SVD approach - both in terms of accuracy and training / testing time.

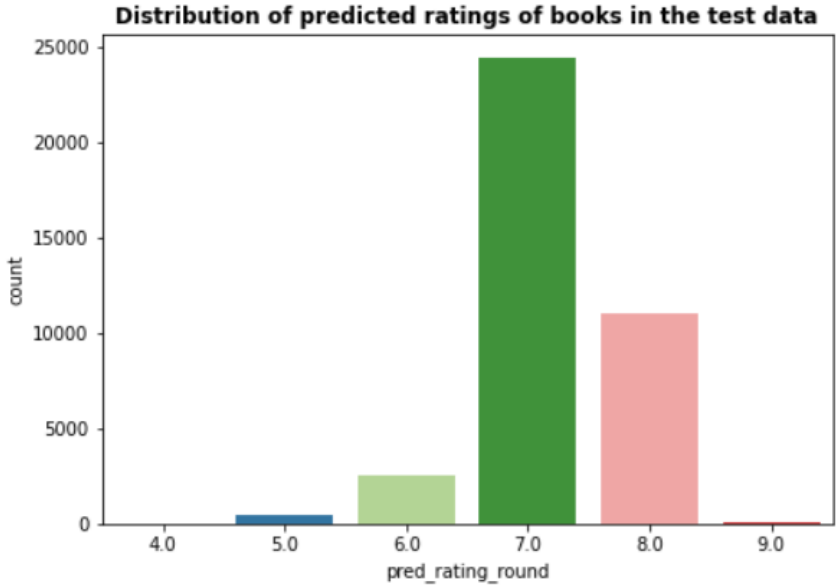
Model Building

SVD with GridSearchCV



Model Building



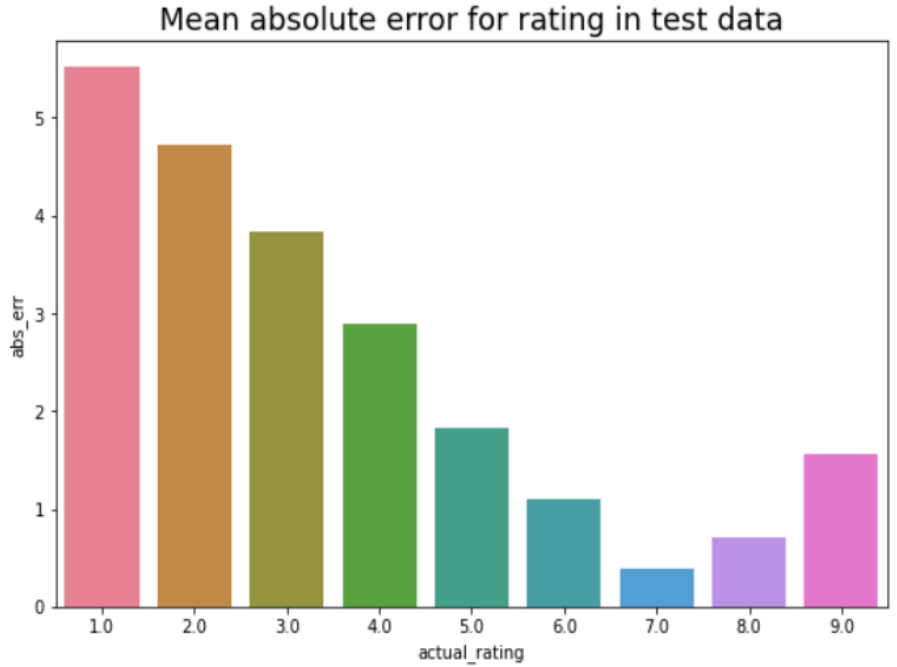


- Most of the users rated the book between 7 and 9. The mode equals 8 but count of ratings 7 and 9 is also noticeable.

- The distribution of predicted ratings in the test set is visibly different compare to actual rating. One more time, 7 is a mode but scores 6 and 9 are clearly less frequent.

Model Building

- The distribution of absolute errors is right-skewed, showing that the majority of errors is lies between 0 and 1.
- There is a long tail that indicates that there are several observations for which the absolute error was close to 7.



- it is observing that absolute error is more than 5 for actual rating 1, with increase in actual rating, absolute error start decreasing linearly upto 7 beyond this value absolute error increases.
- For actual rating of 7, absolute error is minimum.

Model Building

Actual top 5 rated books on test data

	User-ID	ISBN	rating	title	pred_rating
15440	2313	0399146431	5	The Bonesetter's Daughter	7.859308
15456	2313	0812533550	9	Ender's Game (Ender Wiggins Saga (Paperback))	7.721397
15425	2313	0020442602	9	Voyage of the Dawn Treader	7.621987
15446	2313	0553278223	7	The Martian Chronicles	7.343739
15460	2313	0879520388	7	Science and Health with Key to the Scriptures ...	7.296020

Domestic fiction

Science fiction

Child literature

Science fiction

Religion philosophy based on bible

predicted top 5 rated books on test data

	User-ID	ISBN	rating	title	pred_rating
15425	2313	0020442602	9	Voyage of the Dawn Treader	7.621987
15456	2313	0812533550	9	Ender's Game (Ender Wiggins Saga (Paperback))	7.721397
15446	2313	0553278223	7	The Martian Chronicles	7.343739
15460	2313	0879520388	7	Science and Health with Key to the Scriptures ...	7.296020
15440	2313	0399146431	5	The Bonesetter's Daughter	7.859308

- These are the top 5 recommended books for user id 2313
- Both actual and predicted result on test data is almost same.
- As per genres, all the books are related to child.

Model Building

4. Content based filtering

- We have taken only explicit rated data
- We consider 10% of population, i.e. popularity threshold is 52.
- Using Tf vectorizer, we create raw documents to a matrix of TF-IDF features.
- Cosine similarity to find the similarity distance.
- Then we get recommended books

```
Enter book name:-Fahrenheit 451
Recommended Books according to content given in this book are:

Fried Green Tomatoes at the Whistle Stop Cafe
Ender's Game (Ender Wiggins Saga (Paperback))
From the Corner of His Eye
Violets Are Blue
On the Street Where You Live
Message in a Bottle
The Jester
The Bluest Eye
The Summons
The Partner
```

Model Building

5. Memory based collaborative filtering using NearestNeighbours

- We consider only explicit dataset.
- We consider only top 10% frequently rated record with minimum rating 3.
- creating pivot table and replacing null value with zero
- Creating sparse matrix
- Train the model with NearestNeighbours(metric = 'cosine', algorithm = 'brute')
- On the basis of distance we can recommend books

```
Recommendations for Homeport:

1: Divine Evil, with distance of 0.8323123500669327:
2: Sanctuary, with distance of 0.8484176806049264:
3: Morning Glory, with distance of 0.8494120449380568:
4: River's End, with distance of 0.8667295239342745:
5: Irish Rebel (Special Edition, 1328), with distance of 0.8721797423162374:
```

Challenges

- As dataset was very large which led more computation time and every time google colab get crash due to insufficient RAM.
- Missing value imputations and outlier treatment were also difficult.
- Decision on selecting right plot to get required inferences.

Conclusions

- 67.43% users give 0 rating for books and is mostly rated by adult age group.
- 72% of books reader from USA and 9% from Canada
- Stephan King is most popular author and rated by most of the users, Nora Roberts was second after Stephan king.
- The lovely Bones: A Novel is most rated books and it is the most popular books among teens.
- Harry Potter and the Chamber of Secrets (Book 2) is getting maximum weighted score.
- For modelling, it was observed that for model based collaborative filtering SVD technique worked way better than NMF with lower Mean Absolute Error (MAE) and RMSE .
- In cases where a new user or item's rating preference is unknown, collaborative filtering may not be the best recommendation approach. Content-based filtering is preferable.

Thank You