Movie Recommender System

Hizana Nasreen E S

2024-05-10

Contents

1 OVERVIEW

1.1 Introduction and Aim of the Project

The purpose of the project is to create a movie recommendation system, using the MovieLens data set, by building models and evaluating the RMSE for the final Algorithm.

1.2 The MovieLens Data set

The MovieLens data set is a popular open data set for building and evaluating recommender systems. The 10M version of the MovieLens data set is being used for this project. This version contains **approximately 10 million ratings from 71000 users on about 10000 movies**.

Some of the key statistics of the data set are as follows: - The average rating is around 3.5 - The most active user rated over 2,000 movies. - Over 30,000 users rated the most popular movie.

1.3 Methods Used/Followed

1. Library installation and Data Loading
2. Data Summary
3. Data Cleaning and Preprocessing
4. Data Visualization
5. Model Building and Evaluation
    - Mean baseline Model
    - Movie based Model
    - Movie + User based Model
6. Regularization

1.4 About the Model and its RMSE

The model is trained using 90% of the data set(edx set) and tested using the remaining 10% of the data set(final_holdout_test set)

The Root Mean Squared Error(RMSE), is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

The aim of the project is achieved by building a **Regularized Movie+User Model**, which is capable of achieving a RMSE of **0.8628**.

2 METHODS/ANALYSIS

2.1 Library Installation and Data Loading

The project initiates by verifying the presence of essential R packages, including tidyverse, KableExtra, ggplot2, caret. In the event that any of these packages are not installed, the code automatically installs them.

Subsequently, it loads these libraries to facilitate the execution of subsequent data manipulation and analysis tasks. The Movie Lens 10M data set is retrieved and subsequently processed, involving the extraction of ratings and movies data, which forms the foundation of the analysis. The data set is then partitioned into training and testing sets containing 90% of the data and 10% of the data, respectively

2.2 Data Summary

The data set set has **9000055 rows and 6 columns**. Here is a glimpse of the data set.

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi&
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-
6	1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Dra
7	1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fant

userId	movieId	rating	timestamp	title	genres
Min. : 1	Min. : 1	Min. :0.500	Min. :7.897e+08	Length:9000055	Length:9000055
1st Qu.:18124	1st Qu.: 648	1st Qu.:3.000	1st Qu.:9.468e+08	Class :character	Class :character
Median :35738	Median : 1834	Median :4.000	Median :1.035e+09	Mode :character	Mode :character
Mean :35870	Mean : 4122	Mean :3.512	Mean :1.033e+09	NA	NA
3rd Qu.:53607	3rd Qu.: 3626	3rd Qu.:4.000	3rd Qu.:1.127e+09	NA	NA
Max. :71567	Max. :65133	Max. :5.000	Max. :1.231e+09	NA	NA

The attribute **title** and **genre** are of the type **character**, due which the statistics functions(min, 1st Quartile, median, mean, 3rd Quartile and max) are **not applicable**.

Given below is the list of top 25 most frequently rated movies in our data set:

title	count
Pulp Fiction (1994)	31362
Forrest Gump (1994)	31079
Silence of the Lambs, The (1991)	30382
Jurassic Park (1993)	29360
Shawshank Redemption, The (1994)	28015
Braveheart (1995)	26212
Fugitive, The (1993)	25998
Terminator 2: Judgment Day (1991)	25984
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25672
Apollo 13 (1995)	24284
Batman (1989)	24277
Toy Story (1995)	23790
Independence Day (a.k.a. ID4) (1996)	23449
Dances with Wolves (1990)	23367
Schindler's List (1993)	23193
True Lies (1994)	22823
Star Wars: Episode VI - Return of the Jedi (1983)	22584
12 Monkeys (Twelve Monkeys) (1995)	21891
Usual Suspects, The (1995)	21648
Fargo (1996)	21395

Speed (1994)	21361
Aladdin (1992)	21173
Matrix, The (1999)	20908
Star Wars: Episode V - The Empire Strikes Back (1980)	20729
Seven (a.k.a. Se7en) (1995)	20311

2.3 Data Cleaning and Preprocessing

As part of out data quality Analysis, the data is checked for missing values, duplicates, inconsistent titles, and outliers.

The results of the quality check performed have significant impact on our analysis. The presence of outliers, missing values, etc may require additional cleaning and pre-processing to ensure that our results are accurate, accountable and reliable.

Check	Count
Missing Values	0
Duplicates	0
Inconsistent Data	0
Outliers	0

The results of our data quality check implies that, the data is clean and is ready for use.

2.3.1 Extracting the year and month

Extracting the year and month from the timestamp column enables further analysis based on the rating year and month. This can be useful in identifying trends and patterns in the data based on the time of the rating.

Similarly, extracting the release year enables analysis based on the release year, which can be useful in identifying trends and patterns in the data based on the release year.

Before extraction:

	timestamp
1	838985046
2	838983525
4	838983421
5	838983392
6	838983392

After extraction:

	rating_year	rating_month
1	1996	8
2	1996	8
4	1996	8
5	1996	8
6	1996	8

2.3.2 Extracting Release year and Movie titles

Originally, the `title` column in the data set contains the title of the movie along with its release year. Extracting the release years and movie titles enables further analysis based on these variables. This can be useful in identifying trends and patterns in the data based on the release year and movie title.

The cleaning and preprocessing steps performed during this ensure that the data is in a consistent and usable format for further analysis.

Before extraction:

	title
1	Boomerang (1992)
2	Net, The (1995)
4	Outbreak (1995)
5	Stargate (1994)
6	Star Trek: Generations (1994)

After Extraction:

	title	release_year
1	Boomerang	1992
2	Net, The	1995
4	Outbreak	1995
5	Stargate	1994
6	Star Trek: Generations	1994

2.3.3 Extracting Genre

The `genre` column in the data set provided for use holds the genres in a pipe separated format.AS a result of the extraction process, multiple genres separated by pipe characters will be split into separate rows, with each row containing a single genre.

This preprocessing step is essential for subsequent analysis and visualization of the data, as it ensures that missing values are properly handled and that multiple genres are treated as separate categories.

Before extraction:

	genres
1	Comedy|Romance
2	Action|Crime|Thriller
4	Action|Drama|Sci-Fi|Thriller
5	Action|Adventure|Sci-Fi
6	Action|Adventure|Drama|Sci-Fi

After extraction:

genre
Comedy
Romance
Action
Crime
Thriller

2.3.4 Preprocessed Data

After Preprocessing, the unnecessary columns which are not needed further are eliminated. The data given below, after cleaning and preprocessing, is ready for use.

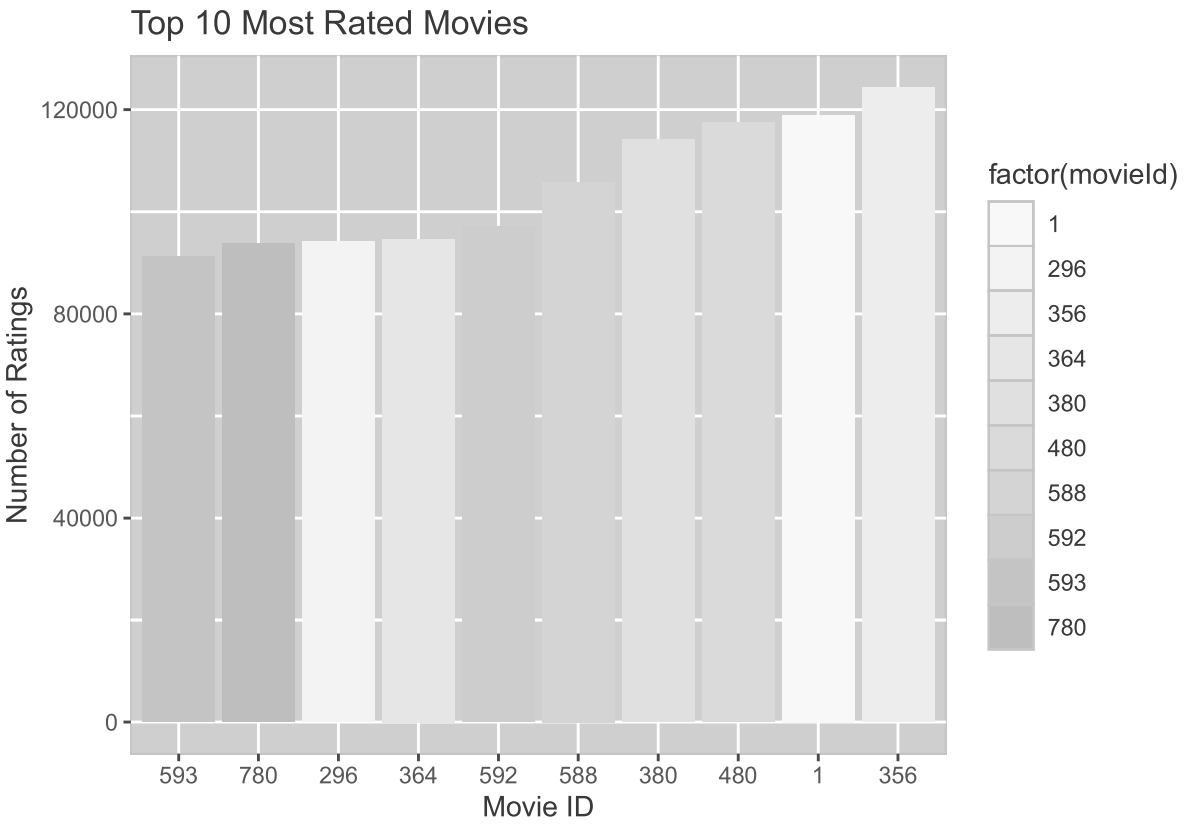
userId	movieId	rating	title	genre	release_year	rating_year	rating_month
1	122	5	Boomerang	Comedy	1992	1996	8
1	122	5	Boomerang	Romance	1992	1996	8
1	185	5	Net, The	Action	1995	1996	8
1	185	5	Net, The	Crime	1995	1996	8
1	185	5	Net, The	Thriller	1995	1996	8
1	292	5	Outbreak	Action	1995	1996	8

2.4 Data Visualization

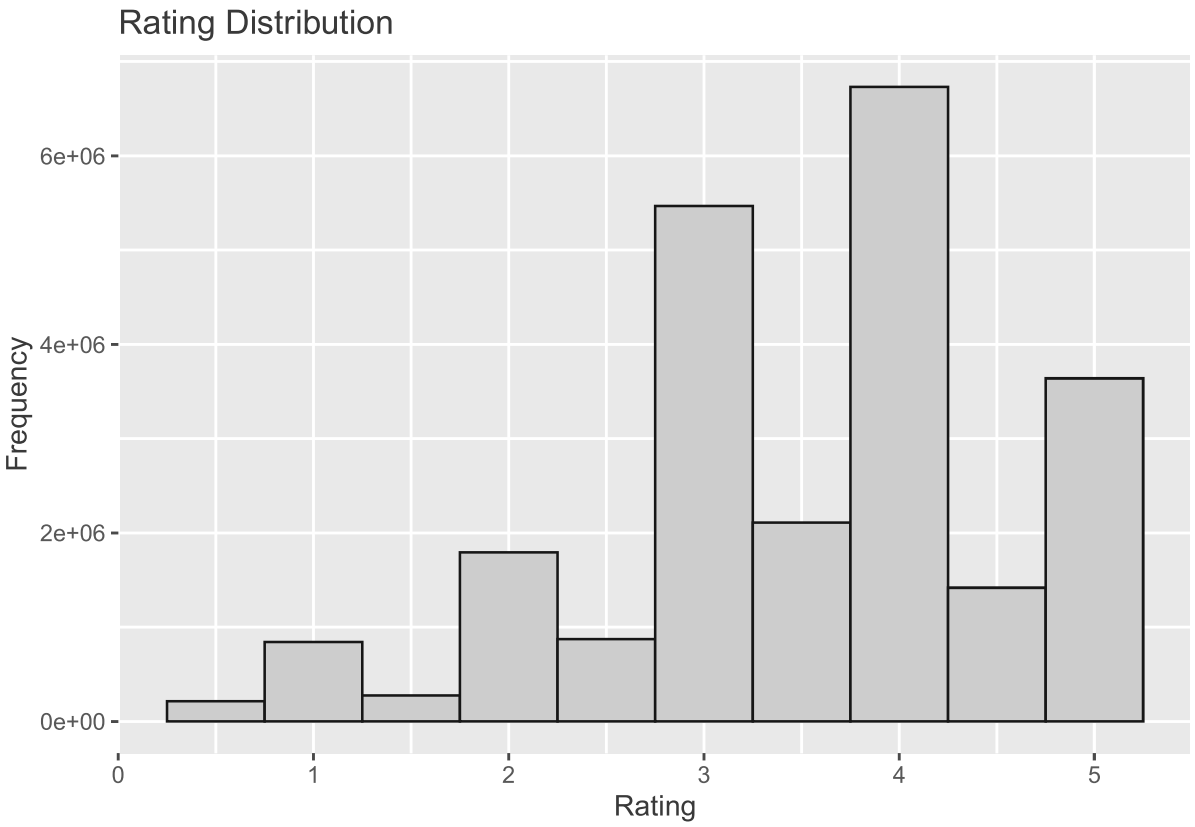
2.4.1 Popularity of Movies

According to the bar plot below, we can observe that, the top 10 most rated movies were released during the years 90s.

```
## Selecting by n_ratings
```

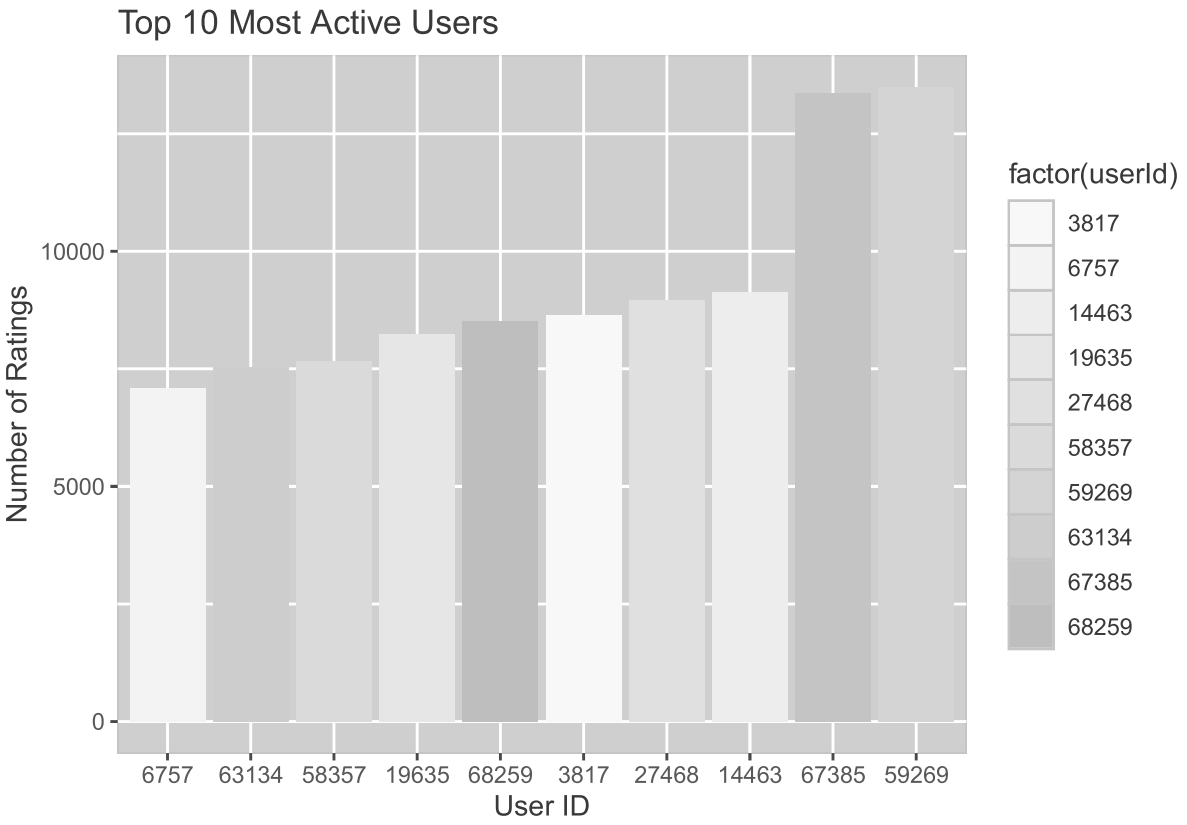


2.4.1.1 Distribution of rating From the distribution of rating among the movies, it is observed that full-star votes were common than half-star votes. The histogram also implies that, the users who disliked, i.e, people who gave a lower rating, were very few compared to other users.

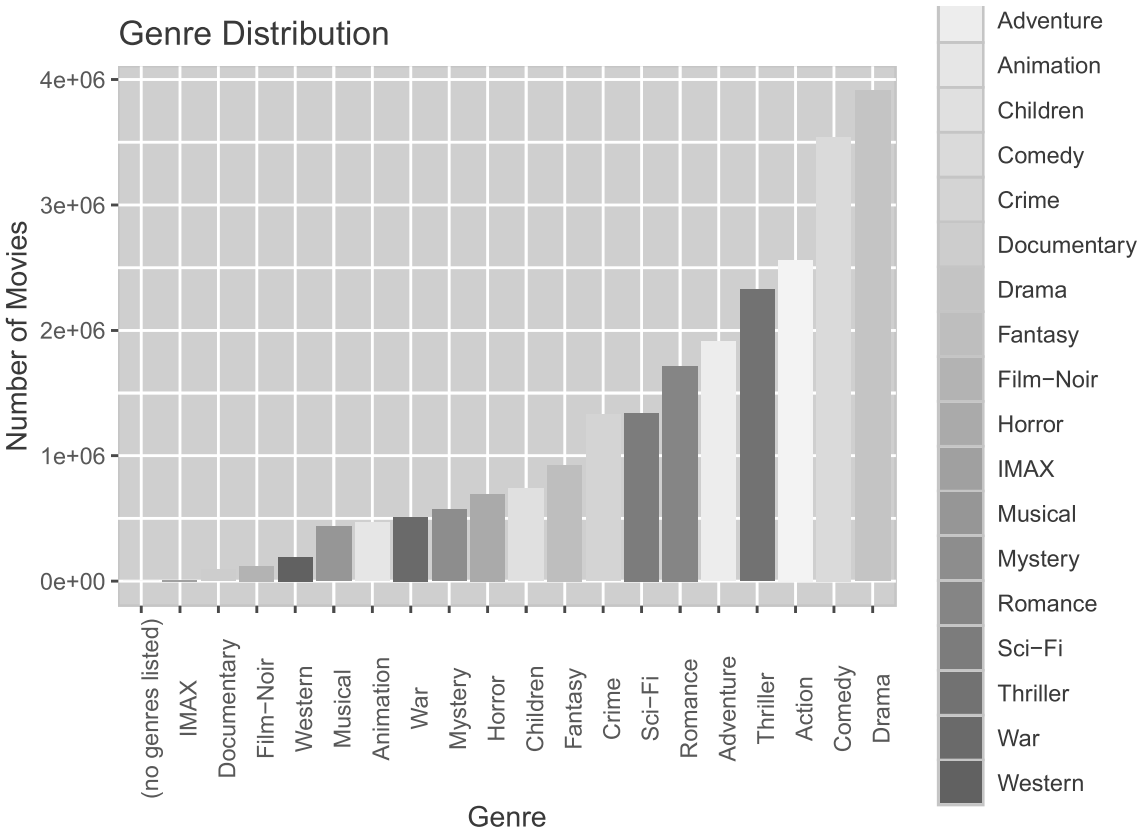


2.4.1.2 User Activity The graph below shows the 10 most active users among the 71000 users. We can observe that, two of the users have given more than 10000 ratings.

```
## Selecting by n_ratings
```

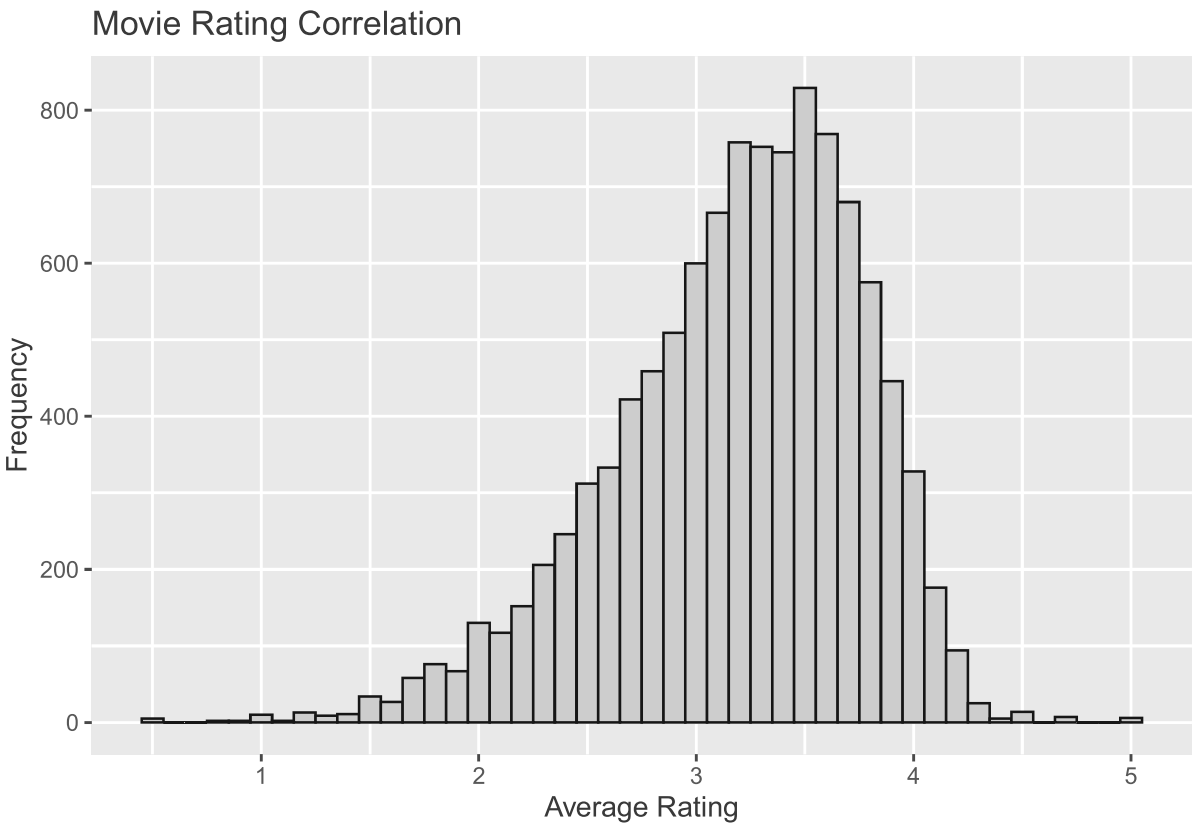


2.4.1.3 Genre Distribution The bar plot below shows how many movies belong to a particular genre, It also give the information about the genres with the most (Drama(3910127)) and least (IMAX(8181)) number of movies .

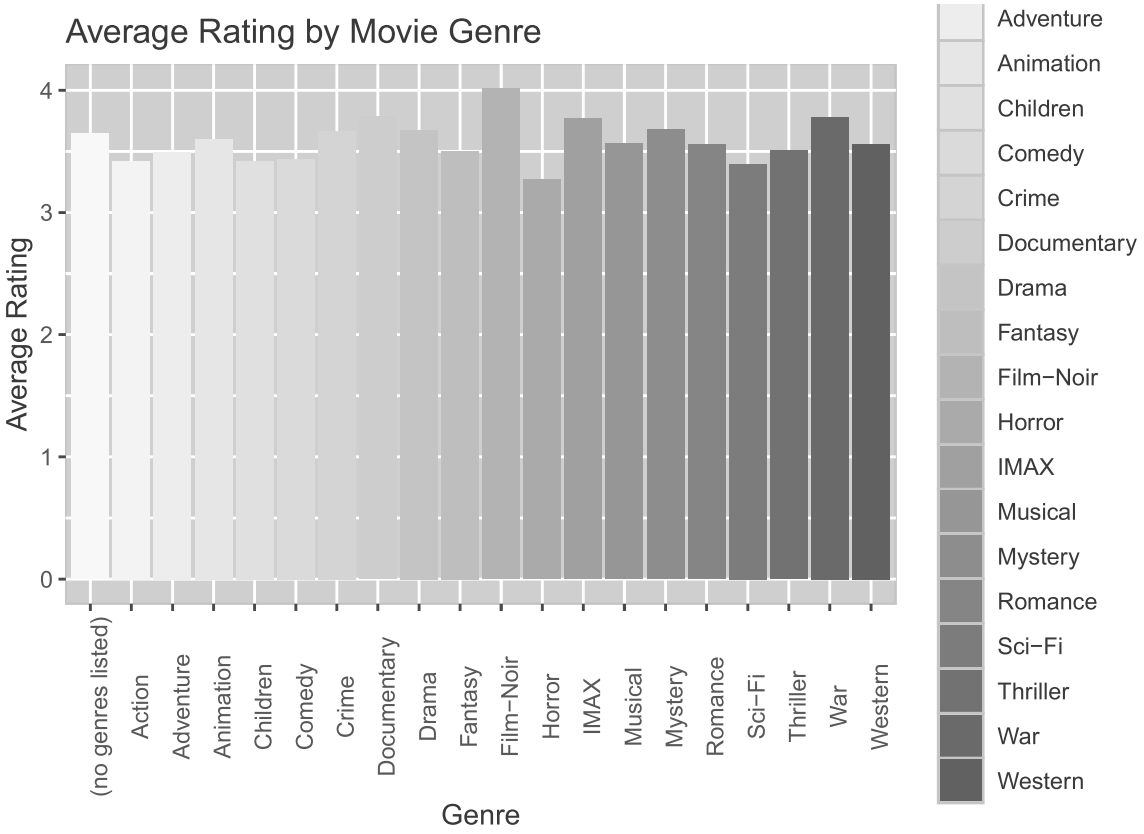


genre	n_movies
Drama	3910127
Comedy	3540930
Action	2560545
Thriller	2325899
Adventure	1908892
Romance	1712100
Sci-Fi	1341183
Crime	1327715
Fantasy	925637
Children	737994
Horror	691485
Mystery	568332
War	511147
Animation	467168
Musical	433080
Western	189394
Film-Noir	118541
Documentary	93066
IMAX	8181
(no genres listed)	7

2.4.1.4 Correlation of the Movie Rating The histogram here visualizes the distribution of average movie ratings in the `edx` data set. We can see that the average rating for the data set is around 3.5.



2.4.1.5 Average Rating by Movie Genre The bar plot shows the average rating per genre. The genre `Film-Noir` has the highest average rating of 4.01 and `Horror` has the least average rating of 3.27.


Average Rating by Movie Genre
Average Rating
4
3
2
1
0
(no genres listed)
Action
Adventure
Animation
Children
Comedy
Crime
Documentary
Drama
Fantasy
Film−Noir
Horror
IMAX
Musical
Mystery
Romance
Sci−Fi
Thriller
War
Western
Genre
Adventure
Animation
Children
Comedy
Crime
Documentary
Drama
Fantasy
Film−Noir
Horror
IMAX
Musical
Mystery
Romance
Sci−Fi
Thriller
War
Western

3 MODEL BUILDING ANG EVALUATION

3.1 Mean Baseline Model

The purpose of this model is to provide a simple baseline model for comparison with more complex models. The mean baseline model assumes that all movies have the same average rating, which is a naive assumption. By comparing the performance of this model with more sophisticated models, we can evaluate the effectiveness of those models in capturing the underlying patterns in the data.

There are certain limitations to this model.

- It assumes that all movies have the same average rating, which is unlikely to be true.
- It does not take into account any individual characteristics of the movies or users.
- It is a very simple model that does not capture any underlying patterns in the data.

Despite these limitations, the mean baseline model provides a useful benchmark for evaluating the performance of more complex models.

The RMSE value obtained on the testing data `final_holdout_test` is found to be **1.052558**

3.2 Movie-based Model

The movie-based model works by assuming that users who have rated a movie similarly in the past will rate it similarly in the future. By calculating the mean rating for each movie, the model can make predictions for new, unseen ratings.

This approach is simple and effective, but it has some limitations, such as:

- It does not take into account individual user preferences.
- It does not capture complex patterns in the data.
- It can be sensitive to outliers and biased ratings.

The RMSE value obtained on the testing data `final_holdout_test` is found to be **0.94107**

3.3 Movie + User based Model

The movie + user-based model works by combining the strengths of both movie-based and user-based models. By accounting for both movie-specific and user-specific effects, the model can capture more complex patterns in the data and provide more accurate predictions. This approach is more effective than the simple movie-based model.

The RMSE value obtained on the testing data `final_holdout_test` is found to be **0.863366**

3.4 RMSE Comparison

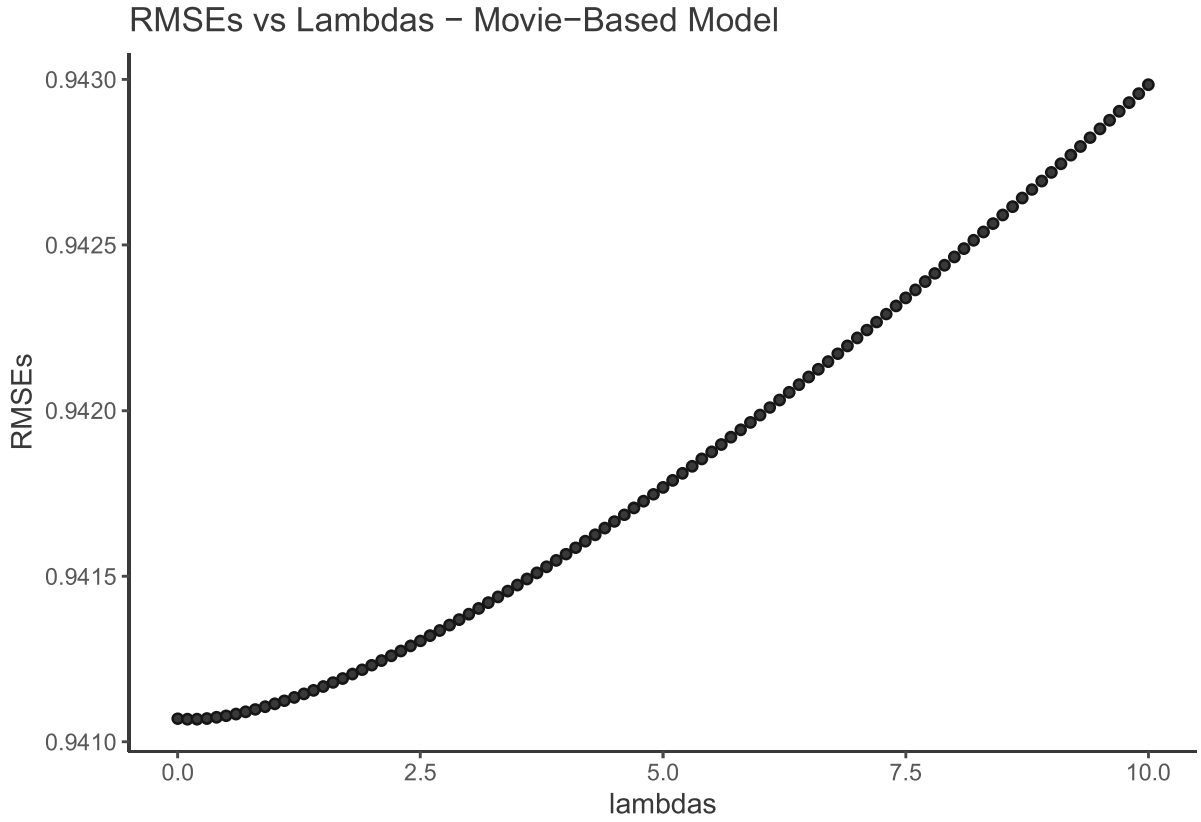
Models	RMSE
Mean Baseline Model RMSE	1.052558
Movie-Based Model RMSE	0.941070
Movie + User-Based Model RMSE	0.863366

4 REGULARIZATION

The goal of hyperparameter tuning is to find the optimal value of the regularization hyperparameter `lambda` that minimizes the Root Mean Squared Error (RMSE) of the movie-based model.

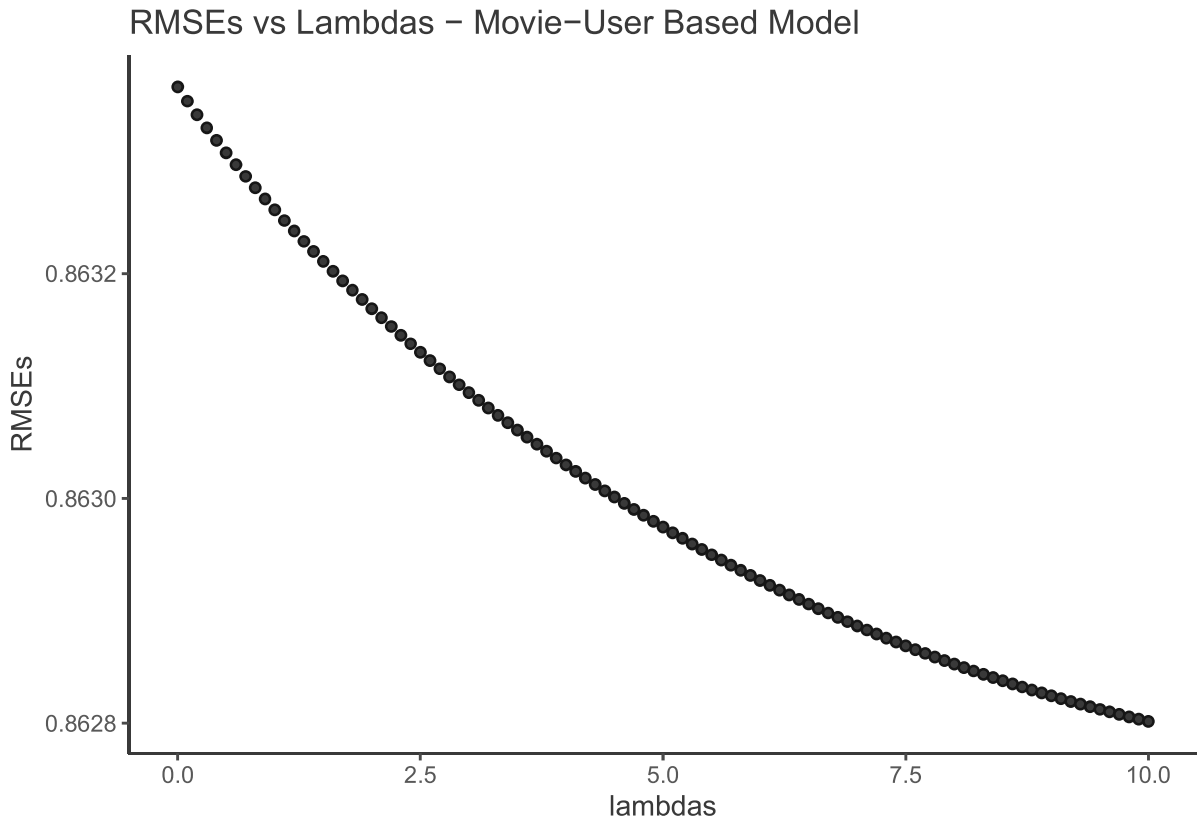
4.1 Regularized Movie-based Model

The Regularized movie + user based model gives an RMSE of **0.9410685** on the `final_holdout_test set`, which not very different from the RMSE obtained on the non-regularized movie-based model.



4.2 Regularized Movie + User based Model

The Regularized movie + user based model gives an RMSE of **0.8628** on the `final_holdout_test set`



4.3 RMSE after Regularization

Regularized_Models	RMSE
Movie-Based Model RMSE	0.9410685
Movie + User-Based Model RMSE	0.8628015

5 RESULTS

The following table compares the RMSE before and after regularization

Before Regularization:

Models	RMSE
Mean Baseline Model RMSE	1.052558
Movie-Based Model RMSE	0.941070
Movie + User-Based Model RMSE	0.863366

After Regularization:

Regularized_Models	RMSE
Movie-Based Model RMSE	0.9410685
Movie + User-Based Model RMSE	0.8628015

6 CONCLUSION

From the above models, we can conclude that `movieId` and `userId` contribute to the training models. The `movie + user based model`, works accurately and gives an RMSE of **0.8628**.

6.1 Future Scope

Further models could be build on the data set by considering the `genre` as a predictor. Additionally potential biases in the models could be investigated and strategies could be developed for mitigating them. The scalability and deployment of the models should also be accounted for.

7 APPENDIX

7.1 EDX Code

```
## [1] "Version : "
```

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## crt            ucrt
## system         x86_64, mingw32
## status
## major          4
## minor          3.1
## year           2023
## month          06
## day            16
## svn rev        84548
## language       R
## version.string R version 4.3.1 (2023-06-16 ucrt)
## nickname       Beagle Scouts
```

8 REFERENCE

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.

10M Version of the MovieLens Dataset

Entire MovieLens Dataset