MovieLens Project

For the Requirement of *HarvardX: PH125.9x Data Science: Capstone Course*

Guler Arsal

September 15, 2024

Abstract

The objective of this project is to develop a movie recommendation system using the MovieLens dataset. The process begins with a detailed examination and preparation of the dataset to ensure it is suitable for analysis. A thorough exploratory analysis follows, incorporating various visualization techniques to uncover key patterns and trends within the data. Based on these insights, the report describes the approach taken to design, train, and evaluate the predictive algorithm. Each iteration of the algorithm is systematically analyzed by comparing it with previous models to assess improvements in performance. The project concludes with an evaluation of the final model, which achieves a Root Mean Square Error (RMSE) of 0.8622 on the validation dataset. The report further discusses the final model's strengths and limitations, offering recommendations for future enhancements.

Contents

List of Tables **3**

List of Figures **4**

1 Introduction **5**

2 Method **5**

2.1 Analysis Tools and Report Compilation 5

2.2 Movielens Dataset 5

2.3 Splitting Datasets 6

3 Results **6**

3.1 Exploratory Analysis 6

3.1.1 Movie Ratings (Outcome Variable) 6

3.1.2 Users 7

3.1.3 Movies 8

3.1.4 Review Date 9

3.1.5 Processing Title Column 12

3.1.6 Movie Title 12

3.1.7 Release Year 14

3.1.8 Movie Genres 15

3.2 Model Development 17

3.2.1 Preparation for Model Development 18

3.2.2 Root Mean Square Error 18

3.2.3 Initial Models (Non-Regularized) 18

3.2.4 Summary of the Initial Models 19

3.2.5 Clamping Adjustments 20

3.2.6 Regularized Models 21

3.3 Final Model Assessment 22

4 Discussion **23**

5 Conclusion **23**

6 References **25**

List of Tables

1 Summary of the Variables in the MovieLens Dataset 5
2 Number of Ratings for Possible Rating Options 6
3 Top 10 Movies Receiving the Most Ratings 13
4 Top 10 Movies Receiving the Highest Ratings 13
5 Genres Ranked Based on Number of Ratings Received 16
6 RMSE Results for the Initial Models 20
7 RMSE Results with and without Clamping 20
8 RMSE Results including the Regularized Model 22

List of Figures

1 Number of Ratings Per User 7
2 Average Rating Per User 8
3 Number of Ratings Per Movie 8
4 Average Rating Per Movie 9
5 Number of Ratings by Year Reviewed 10
6 Average Rating by Week Reviewed 10
7 Distribution of Average Rating by Year Reviewed 11
8 Number of Ratings for Possible Rating Options Before and After 2003 12
9 Average Rating by Release Year 14
10 Number of Ratings by Release Year 15
11 Genre Ratings (Ranked by Mean Rating) 17
12 RMSE Against Lambda 21
13 Histogram of Residuals 23

1 Introduction

The objective of this paper is to develop a movie recommendation system using the MovieLens dataset. Such a system aims to suggest films to users by analyzing factors like viewing history and behavior, striving to predict movies that users will likely enjoy. By automating the decision-making process, these systems help users make informed choices about what to watch (Ricci et al., 2011). They are widely employed by streaming platforms like Netflix and Amazon Prime, enhancing user experience through personalized recommendations.

2 Method

2.1 Analysis Tools and Report Compilation

All analyses were conducted using R version 4.3.2, a comprehensive software environment developed by the R Foundation for Statistical Computing, renowned for its robust data analysis and visualization capabilities (R Core Team, 2023). The report was compiled using R Markdown within RStudio, an integrated development environment specifically designed for R programming (RStudio Team, 2023). Both R and RStudio are open-source applications freely available to the public.

2.2 Movielens Dataset

The *MovieLens* dataset, which is publicly accessible at http://files.grouplens.org/datasets/movielens/ml-10m.zip, was acquired using the course-provided code. It contains 10,000,054 rows, each representing a rating given by a user to a specific movie. Table 1 provides a detailed description of the variables included in the dataset.

Table 1
Summary of the Variables in the MovieLens Dataset

Variable Name	Data Type	Description
userId	integer	Unique identifier assigned to each user (n = 69,878).
movieId	integer	Unique identifier assigned to each movie (n = 10,677).
rating	numeric	Outcome variable to be predicted. Ratings are on a scale of 0.5 stars (worst rating) to 5 stars (best rating), in increments of 0.5.
timestamp	integer	Time at which the user provided the rating. It represents seconds since January 1, 1970.
title	character	Name of the movie, including the release year in parenthesis.
genres	character	Genres associated with a particular movie, such as comedy, horror, and drama. It includes every genre that applies to the movie. As movies could be classified to one or more genres, there is 797 unique genre combinations listed in this column.

2.3 Splitting Datasets

The MovieLens dataset was partitioned using the course-provided code into a training set (*edx*, 90% of data) and an evaluation set (*final_holdout_test*, 10% of data). The edx dataset (9,000,061 rows), which was used to develop machine learning models, includes 69,878 unique users and 10,677 unique movies. The edx dataset was further split into two sets by 80% and 20%, labelled as *train_set* and *test_set*, respectively. The *semi_join* function ensured the test set only included users and movies present in the train set, while *anti_join* added the removed data back into the train set to maximize training data.

3 Results

3.1 Exploratory Analysis

Data exploration is crucial for model building (Tukey, 1977). It involves analyzing and understanding the data to guide decision-making. The edx dataset variables were explored through techniques like examining data distribution, visualizing and summarizing data, and transforming variables.

3.1.1 Movie Ratings (Outcome Variable)

When rating movies, users can select a value ranging from 0.5 stars (the lowest rating) to 5 stars (the highest rating) in increments of 0.5, resulting in a total of ten possible rating options. Table 2 shows occurrences of each rating options in the edx dataset. The rating of 4.0 was the most common, while a rating of 0.5 was the least common options. This pattern suggests a tendency toward positive ratings. The overall mean rating of **3.51** further supports this observation.

Table 2
Number of Ratings for Possible Rating Options

Rating Options	Number of Ratings
0.5	85,420
1.0	345,935
1.5	106,379
2.0	710,998
2.5	332,783
3.0	2,121,638
3.5	792,037
4.0	2,588,021
4.5	526,309
5.0	1,390,541

The examination of Table 2 also demonstrated that users exhibited a preference for whole number ratings (1.0, 2.0, 3.0, 4.0, 5.0) over decimal values (0.5, 1.5, 2.5, 3.5, 4.5). The users' preference for round numbers when making judgments aligns with the broader literature on human cognitive biases. This tendency is consistent with anchoring bias, where people rely on familiar or easily

accessible reference points to simplify decision-making.(Tversky & Kahneman, 1974) and can be linked to the brain's efficient coding strategy (Prat-Carrabin & Woodford, 2022).

3.1.2 Users

Figure 1 displays a histogram of users and their respective rating counts on a log scale, demonstrating a pronounced positive skew. This skew reflects the presence of extreme values on the right side of the distribution. While the average user has rated 129 movies, as indicated by the red line, the median user has only rated 62 movies, marked by the blue line. A notable feature of the distribution is the right tail, which includes a subset of highly active users—612—who have rated more than 1,000 movies, as indicated by the green line. These users represent approximately 0.9% of the total user base.

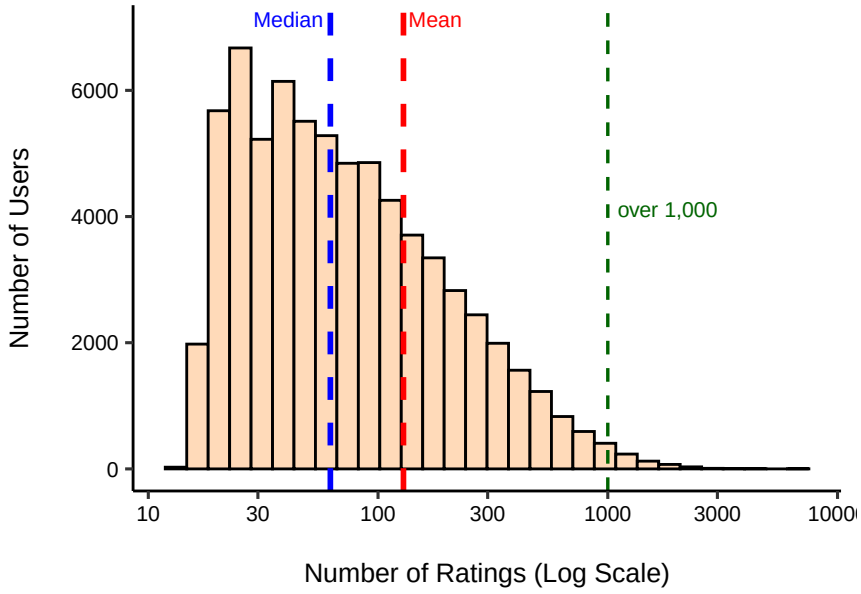

Figure 1: Number of Ratings Per User

Figure 2 shows the histogram of average rating by number of user. It exhibits a symmetrical, bell-shaped curve indicating that the underlying data is normally distributed. The histogram shows that users generally assigned high average ratings, with a mean of 3.61, reflecting a selection bias toward positive ratings. This bias can be explained in the following way. Users generally rate movies only after watching them to completion, and they typically begin viewing films they expect to enjoy. Consequently, the set of rated movies is biased toward those that initially piqued the viewer's interest, thus increasing the likelihood of favorable evaluations.

It is important to note that there is considerable variability among users. Some users tend to be more generous, providing higher ratings, whereas others are more critical, reflected in their relatively lower average ratings. The variability among users suggests that incorporating a user effect could enhance the recommendation system's accuracy.

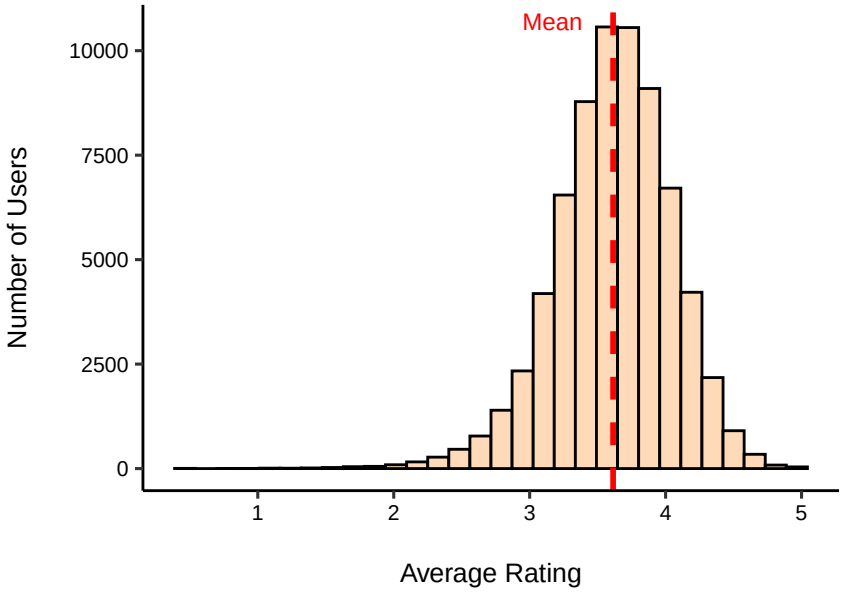

Figure 2: Average Rating Per User

3.1.3 Movies

Figure 3 depicts the distribution of movies by number of ratings on a log scale. The histogram is is positively skewed. A movie on average received 843 reviews, whereas the median movie only received 122 reviews. The plot shows that certain movies were more popular and received ratings more frequently than others. Some movies received very few ratings, with 121 movies rated only once.

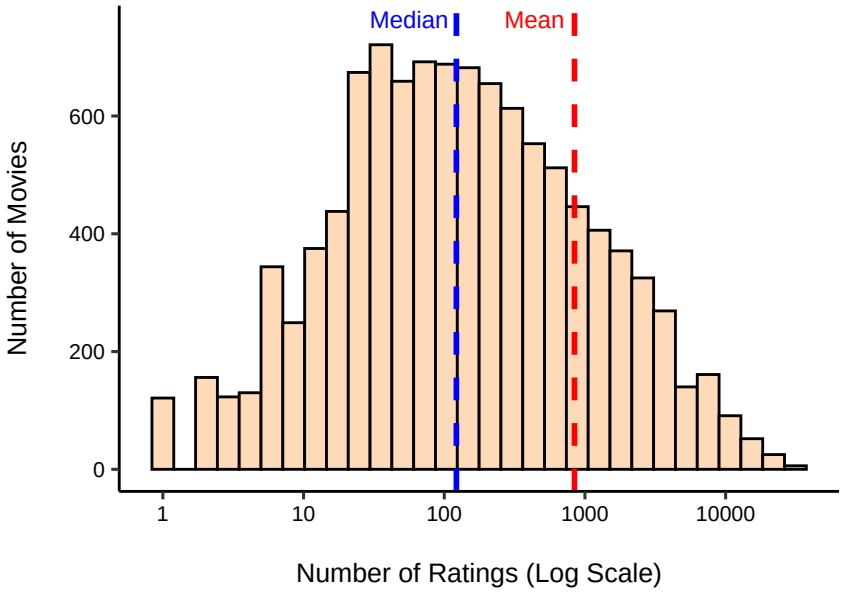

Figure 3: Number of Ratings Per Movie

Figure 4 depicts the distribution of movies by average rating. The histogram is very slightly skewed to the left. Movies were generally assigned with high average ratings, with a mean of 3.19, and a median of 3.26. These findings demonstrate that including a movie effect in the training algorithm could improve recommendation accuracy.

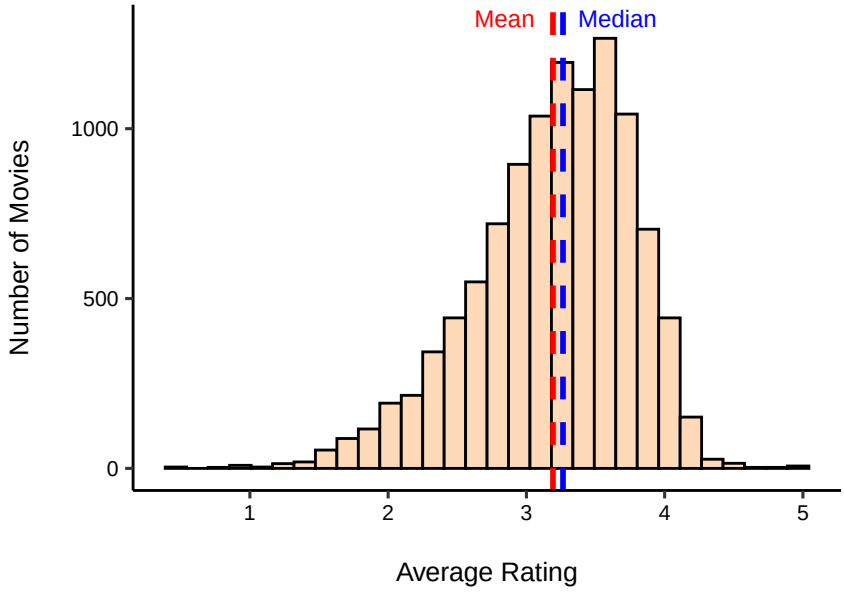

Figure 4: Average Rating Per Movie

3.1.4 Review Date

The *timestamp* variable is represented as the number of seconds elapsed since January 1, 1970, known as the Unix epoch. For example, a timestamp value of 946684800 corresponds to January 1, 2000. To enable time-based analysis of the reviews, the timestamp was first converted to a date format, with the time component removed. Two key transformations were then applied:

1. A new column was created to represent the week of each review, with dates rounded to the nearest week.
2. Another column was generated to capture only the year of each review (e.g., "2020").

Figure 5 illustrates the distribution of the number of ratings across each year. The earliest recorded review dates back to 1995, with the most recent occurring in 2009. Notably, only two ratings were submitted in 1995, whereas both 2000 and 2005 saw over one million ratings each.

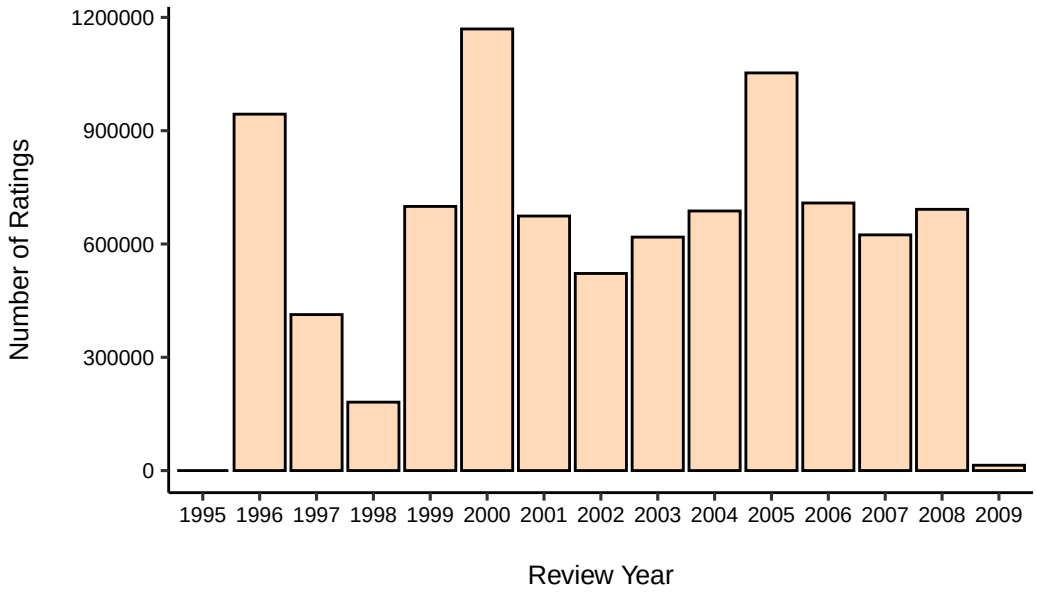

Figure 5: Number of Ratings by Year Reviewed

Figure 6 presents a time series analysis of average rating across review dates, with data analyzed at a weekly granularity rather than annually. The average rating remain relatively stable initially, followed by a slight decline during the late 1990s and early 2000s. After 2005, a gradual increase in average rating is observed.

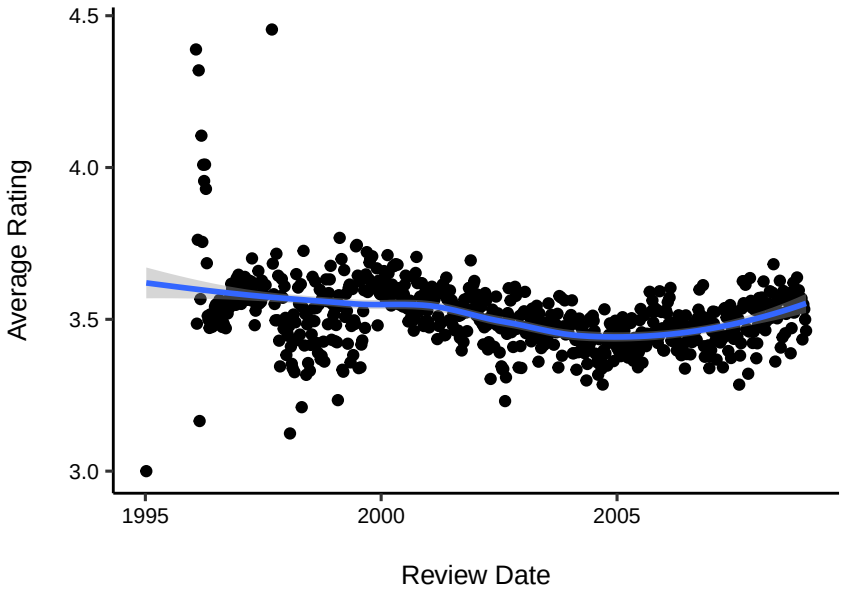

Figure 6: Average Rating by Week Reviewed

To gain a better understanding of the overall trend of average rating, a boxplot visualizing the distribution of movie ratings across different review years was also generated (see Figure 7). In this figure, the medians are represented by the bold horizontal lines within the boxes and the outliers

are indicated by the dots outside the whiskers. Notably, films reviewed before 2003 predominantly show median ratings of 3 or 4, with outliers scoring 1. Conversely, movies reviewed after 2003 generally exhibit median ratings around 3.5, with outlier scores of 1 and 0.5. This finding may suggest the introduction of new rating options in 2023.

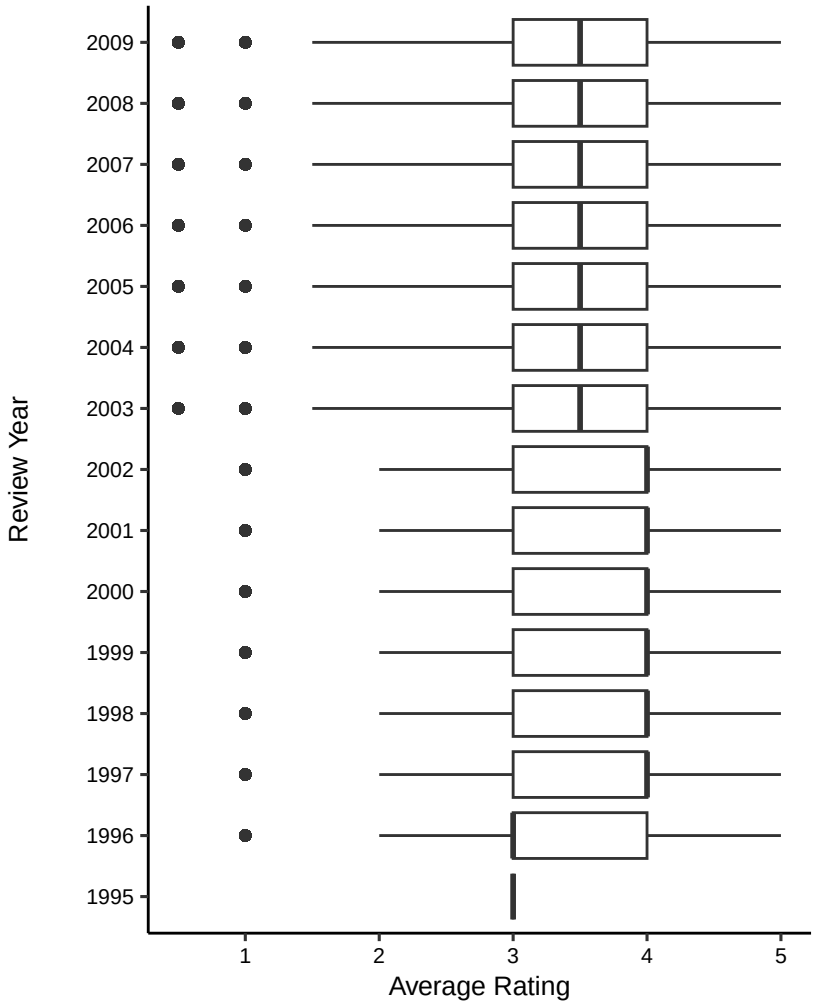

Figure 7: Distribution of Average Rating by Year Reviewed

Figure 8 below provides evidence for the introduction of new rating options in 2003. It shows the number of ratings for each rating option, separated by whether the review year is before or after 2003. In the left panel of the figure, which represents the period before 2003, it is evident that users could only assign whole-number ratings (1, 2, 3, 4, or 5), with no decimal-number ratings present. However, starting in 2003, the rating system was revised to include half-point increments, allowing for more refined user evaluations.

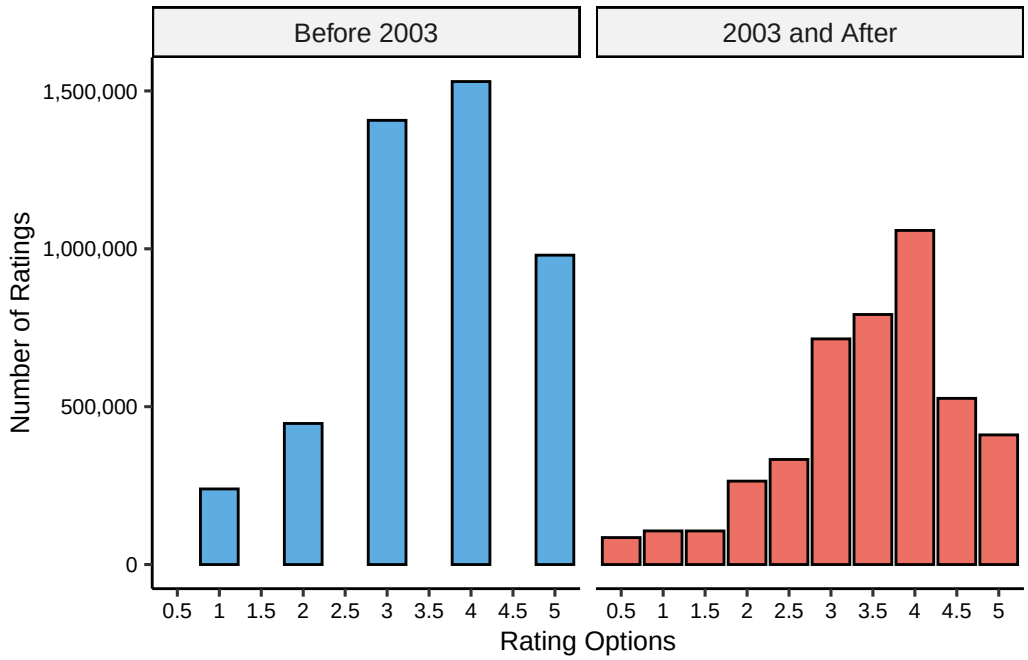

Figure 8: Number of Ratings for Possible Rating Options Before and After 2003

3.1.5 Processing Title Column

A combination of string manipulation functions were used to clean and process the *title* column to extract the release year and adjust the title formatting. First, any leading or trailing whitespace from the title column was removed using the *str_trim* function from the *stringr* package. Second, the *extract* function was used to split the title column into two new columns. One column captured the movie title. The other column captured the year or years (sometimes a range) that are in the title. Afterwards, the release year was converted to an integer. If the extracted release year contains more than four characters (due to the existence of ranges for some), the first year in the range is selected using *str_split* function.

3.1.6 Movie Title

Table 3 lists the top 10 movies receiving the most ratings. Leading the list is *Pulp Fiction*, which has received 31,336 ratings. Following closely is *Forrest Gump* with 31,076 ratings, and *The Silence of the Lambs* ranks third with 30,280 ratings. Other films in the top 10 include *Jurassic Park, The Shawshank Redemption, Braveheart, Terminator 2: Judgment Day, The Fugitive, Star Wars: Episode IV - A New Hope*, and *Batman*. These movies have garnered a substantial number of ratings, reflecting their widespread popularity and appeal. Additionally, these movies received high average ratings, ranging from 3.39 for *Batman* to 4.46 for *The Shawshank Redemption*.

Table 3
Top 10 Movies Receiving the Most Ratings

Rank	Title	Frequency	Average
1	Pulp Fiction	31,336	4.16
2	Forrest Gump	31,076	4.01
3	Silence of the Lambs, The	30,280	4.21
4	Jurassic Park	29,291	3.66
5	Shawshank Redemption, The	27,988	4.46
6	Braveheart	26,258	4.08
7	Terminator 2: Judgment Day	26,115	3.93
8	Fugitive, The	26,070	4.01
9	Star Wars: Episode IV - A New Hope (a.k.a. Star Wars)	25,809	4.22
10	Batman	24,664	3.39

Table 4 displays the top 10 highest-rated movies, all of which have exceptionally high average ratings ranging from 4.75 to a perfect score of 5.00. However, these films have been rated by only a few users. Notably, all but one of the movies with an average rating of 5.00 have been rated only once. Due to these small sample sizes, the ratings are not reliable and may result in noisy estimates. Therefore, these ratings should be interpreted with caution.

Table 4
Top 10 Movies Receiving the Highest Ratings

Rank	Title	Average	Frequency
1	Blue Light, The (Das Blaue Licht)	5.00	1
2	Constantine's Sword	5.00	1
3	Fighting Elegy (Kenka erejii)	5.00	1
4	Hellhounds on My Trail	5.00	1
5	Satan's Tango (Sátántangó)	5.00	2
6	Shadows of Forgotten Ancestors	5.00	1
7	Sun Alley (Sonnenallee)	5.00	1
8	Human Condition II, The (Ningen no joken II)	4.83	3
9	Human Condition III, The (Ningen no joken III)	4.75	4
10	Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva)	4.75	4

Note: The small sample sizes make these ratings unreliable

3.1.7 Release Year

Figure 9 presents a scatter plot illustrating the trend in average movie ratings over time, categorized by the year of release. The plot features a smooth blue line, accompanied by a shaded confidence interval, indicating that average ratings peaked for movies released between 1940 and 1950 and have generally declined for films released in subsequent decades. This peak may suggest that viewers rate movies from the 1940s and 1950s with a sense of nostalgia, reflecting their enduring popularity and cultural significance. It is also possible that the decline in movie ratings over the past fifty years could be attributed to the fact that more recent movies have had less time for users to rate them.

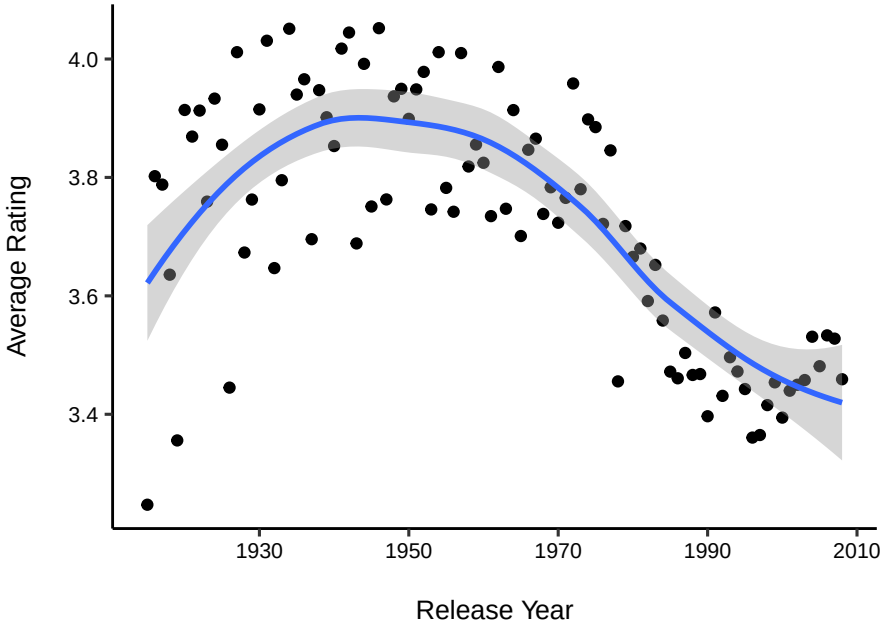

Figure 9: Average Rating by Release Year

To further investigate the hypotheses mentioned in the previous paragraph, a line plot displaying the number of movie ratings by release year was created (Figure 10). The plot shows a relatively low number of ratings for movies released before 1970. There is a gradual increase in the number of ratings for movies released from the 1970s to the 1980s. Most ratings are concentrated on films released in the 1990s, with a noticeable peak in 1995. This distribution challenges the hypothesis that older films would attract significant viewer engagement. It also highlights that point estimates for older films (with fewer ratings) could introduce variability, affecting the reliability of any predictive models based on this dataset.

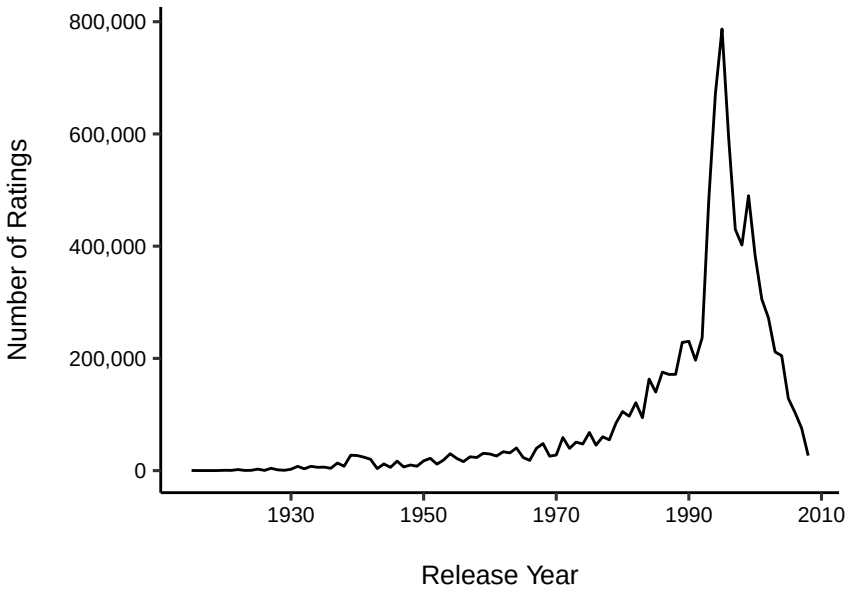

Figure 10: Number of Ratings by Release Year

3.1.8 Movie Genres

Recall that within the dataset, some movies were assigned to multiple genres, resulting in a total of 797 unique genre combinations. To facilitate analysis, the *genres* column was restructured, transforming each genre associated with a movie into a distinct row. For instance, a movie classified under "Action|Comedy" would be represented by two separate rows: one for "Action" and another for "Comedy." This approach allowed for the isolation of individual genres and enabled the ranking of distinct genre categories.

Table 5 provides a comprehensive overview of different movie genres, ranked by the total number of ratings they received. The *Drama* genre garnered the highest number of ratings, followed by the *Comedy* and *Action* genres. These genres attract the most viewer engagement and are widely watched and rated. In contrast, the *IMAX* genre received the fewest ratings, excluding six ratings that did not correspond to any genre for unknown reasons.

Table 5 also highlights which genres tend to have higher or lower average ratings, indicating perceived quality. Genres such as *Film-Noir*, *Documentary*, and *War* have higher average ratings, indicating that movies in these genres are generally well-received by viewers. Conversely, genres like *Horror* and *Sci-Fi* have lower average ratings, suggesting lower satisfaction levels among viewers.

Table 5 also shows that some genres, like *Drama* and *Comedy*, have a large number of movies, suggesting a broad selection within these categories. In contrast, genres such as *IMAX* and *Film-Noir* have fewer movies available, indicating a more limited selection.

Table 5
Genres Ranked Based on Number of Ratings Received

Rank	Genres	Number of Ratings	Average Rating	Number of Movies
1	Drama	3,909,401	3.67	5,336
2	Comedy	3,541,284	3.44	3,703
3	Action	2,560,649	3.42	1,473
4	Thriller	2,325,349	3.51	1,705
5	Adventure	1,908,692	3.49	1,025
6	Romance	1,712,232	3.55	1,685
7	Sci-Fi	1,341,750	3.40	754
8	Crime	1,326,917	3.67	1,117
9	Fantasy	925,624	3.50	543
10	Children	737,851	3.42	528
11	Horror	691,407	3.27	1,013
12	Mystery	567,865	3.68	509
13	War	511,330	3.78	510
14	Animation	467,220	3.60	286
15	Musical	432,960	3.56	436
16	Western	189,234	3.56	275
17	Film-Noir	118,394	4.01	148
18	Documentary	93,252	3.78	481
19	IMAX	8,190	3.76	29
20	(no genres listed)	6	3.50	1

Figure 11 below displays boxplots showing the distribution of average ratings for different movie genres. The boxplots reveal that all genres generally received positive evaluations, with both the median ratings (represented by the horizontal lines inside the boxes) and the mean ratings (depicted by the red dots) consistently above the midpoint of 3.

As mentioned previously, the *Film-Noir* genre has the highest average rating, with a mean of 4.01 and a median of 4. On the other hand, the *Horror* genre has the lowest average rating, with a mean of 3.27 and a median of 3.5. However, please note that the whiskers of the boxplots, which extend from each quartile to the minimum and maximum values, show that the *Film-Noir* genre has the largest interquartile range. This greater variability in ratings for the *Film-Noir* genre likely reflects its relatively small sample size of ratings (n = 118,394), which can result in more pronounced fluctuations and spread in the data.

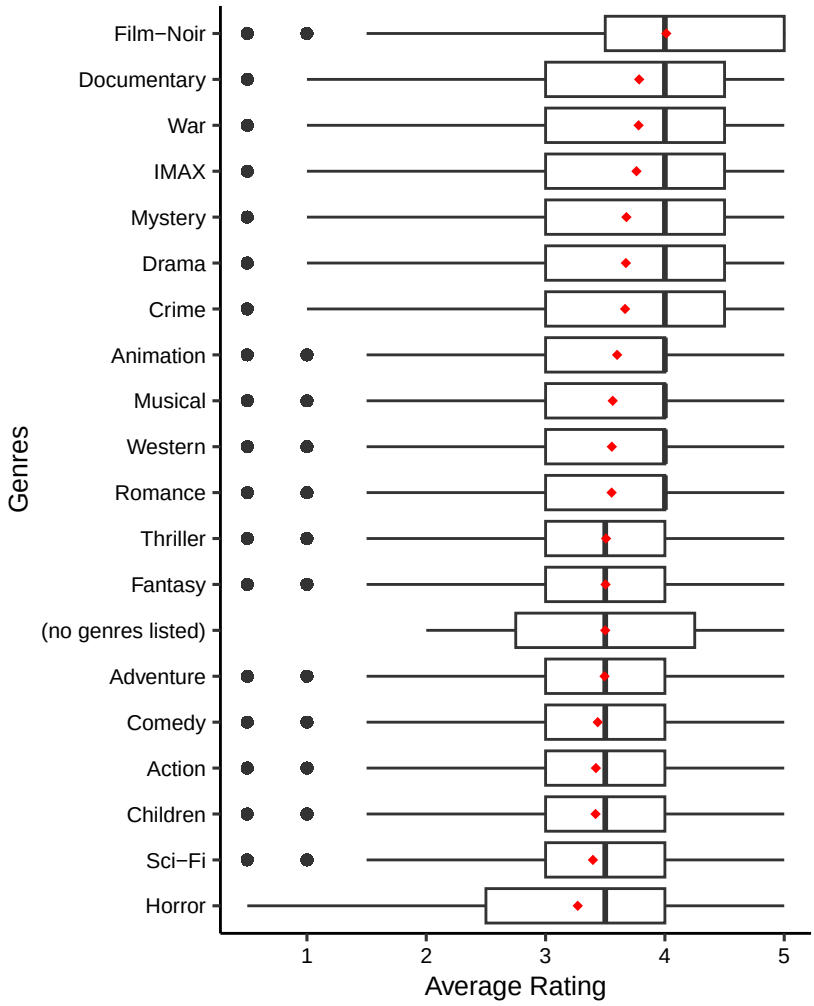

Figure 11: Genre Ratings (Ranked by Mean Rating)

3.2 Model Development

The exploratory data analysis provided valuable insights that are crucial for developing an algorithm to estimate movie ratings. Key findings from this analysis, such as user behavior patterns, the impact of genre and release year on ratings will be integrated into the upcoming machine learning models to enhance predictive performance. In the model development process, the movie effect will be introduced first, as it is expected to play a significant role in predicting ratings. This will be followed by the addition of the user effect, which is also anticipated to have substantial influence. These two effects, being the most important, will form the foundation of the model. Subsequently, other features, including genre, release year, and review date, will be added sequentially to further improve model accuracy.

3.2.1 Preparation for Model Development

Before building the models, both the *train_set* and *test_set* datasets were updated to incorporate the modifications made to the edx dataset. This step ensures that the data used for training and evaluation is aligned with the exploratory data analysis and accurately reflects the key factors identified as important for predicting movie ratings. By ensuring consistency across datasets, the models are better positioned to capture the effects of these critical variables.

3.2.2 Root Mean Square Error

The Root Mean Square Error (RMSE) will be employed to evaluate and compare the performance of different models in accurately predicting movie ratings. RMSE is a widely used metric for assessing the predictive accuracy of a model when dealing with quantitative data. It is defined as "the square root of the average of squared differences between predicted and observed values, which provides an indication of how well the model predictions match the actual data" (Kutner et al., 2005, p. 248).

The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

- n is the number of observations,
- y_i represents the actual observed value,
- $\hat{y}_i$ represents the predicted value.

When comparing models, a lower RMSE value indicates that the predicted ratings are closer to the actual observed ratings, thereby reflecting a more accurate and better-performing model. In contrast, a higher RMSE value indicates less accurate predictions.

3.2.3 Initial Models (Non-Regularized)

Model 1: Baseline Model

The first model is a basic movie recommendation system that naively predicts the same rating for all movies, regardless of individual user preferences. This baseline model serves as a reference point for assessing the effectiveness of more advanced models developed in subsequent sections. The Baseline Model yielded an RMSE of 1.0596.

Model 2: Movie Effect Model

The second model, referred to as the Movie Effect Model, adjusts the global average rating to account for individual movie biases. This adjustment improves the model's ability to predict ratings, resulting in an RMSE of 0.9432. Incorporating the movie effect significantly enhances the model's predictive accuracy compared to using the global average alone.

Model 3: User Effect Model

To account for individual users' tendencies to rate movies higher or lower than average, the third model, known as the User Effect Model, was developed. In this model, predictions are generated by adjusting the global average rating to reflect both the specific biases of individual movies and the unique tendencies of individual users. The inclusion of the user effect further improves predictive accuracy, with an RMSE of 0.8655.

Model 4: Genre Effect Model

The fourth model builds upon its predecessors by adjusting for user, movie, and genre effects to predict movie ratings. This model, referred to as the Genre Effect model, has two versions.

a. Model 4a treats genres as a single, unified category, without distinguishing between sub-genres. This approach yielded an RMSE of 0.8652.

b. To enhance predictive accuracy, Model 4b: Genre-Specific Effect Model was developed. This version separates genres into individual sub-genres, allowing the model to capture the distinct influence of each sub-genre on movie ratings. This refinement improved performance, yielding an RMSE of of 0.864.

Model 5: Release Year Effect Model

The fifth model introduces the Release Year Effect, recognizing that the cultural, social, and historical context at the time of a movie's release can influence audience ratings. Incorporating the release year effect further improves the model's predictive accuracy, resulting in an RMSE of 0.8637.

Model 6: Review Date Effect Model

The sixth model, referred to as the Review Date Effect Model, incorporates the date at which each rating was provided, acknowledging that rating patterns can evolve over time due to various factors. This model achieved the lowest RMSE among the six models, with a value of 0.8635.

3.2.4 Summary of the Initial Models

Table 6 presents the RMSE values for the six models developed in this section, each progressively incorporating additional effects to improve predictive accuracy. As the models become more complex, the RMSE decreases, indicating that each added feature contributes to more accurate predictions.

The table also includes a column showing the difference in RMSE between consecutive models. A negative value indicates an improvement in model performance, with larger negative numbers reflecting more significant gains. The largest improvement in RMSE occurs when adding the Movie Effect in Model 2. While features like release year and review date help fine-tune the model, their contribution is minimal compared to the substantial gains achieved by adding movie and user information.

Table 6
RMSE Results for the Initial Models

Model	Method	RMSE	Diff
Model 1	Average Rating	1.0596	NA
Model 2	Movie Effect	0.9432	-0.1164
Model 3	Movie + User Effects	0.8655	-0.0777
Model 4a	Movie + User + Genre Effects	0.8652	-0.0003
Model 4b	Movie + User + Genre-Specific Effects	0.8640	-0.0012
Model 5	Movie + User + Genre-Specific + Release Year Effects	0.8637	-0.0003
Model 6	Movie + User + Genre-Specific + Release Year + Review Date Effects	0.8635	-0.0002

3.2.5 Clamping Adjustments

The clamp function was introduced to enhance the accuracy of RMSE calculations by ensuring that predicted movie ratings stayed within the valid range, which spans from 0.5 (the lowest possible rating) to 5 (the highest possible rating). In the initial models, some predictions fell outside this acceptable range, which could have distorted the RMSE. By applying the clamp function, these out-of-bounds predictions were adjusted accordingly, maintaining predictions within the realistic rating scale and improving the overall accuracy of the model (Goodfellow, Bengio, & Courville, 2016).

Table 7 summarizes RMSE results for the initial six models with and without clamping technique. As expected, the RMSE values are marginally improved after clamping. This indicates that restricting predictions to a valid range can slightly enhance model accuracy.

Table 7
RMSE Results with and without Clamping

Model	Method	RMSE	Clamp
Model 1	Average Rating	1.0596	1.0596
Model 2	Movie Effect	0.9432	0.9432
Model 3	Movie + User Effects	0.8655	0.8653
Model 4a	Movie + User + Genre Effects	0.8652	0.8650
Model 4b	Movie + User + Genre-Specific Effects	0.8640	0.8638
Model 5	Movie + User + Genre-Specific + Release Year Effects	0.8637	0.8635
Model 6	Movie + User + Genre-Specific + Release Year + Review Date Effects	0.8635	0.8633

3.2.6 Regularized Models

An additional model was developed to address the potential for unreliable predictions caused by sparse data points observed during exploratory data analysis. Specifically, some movies received very few ratings (e.g., only one rating per movie), and certain users exhibited disproportionately high levels of activity. These factors, if left unaddressed, could lead to overfitting, where the model gives undue influence to data with limited observations. To mitigate this risk, regularization techniques were applied, as recommended by Hastie et al. (2009) and Irizarry (2019).

Model 7: Regularized Model
This model was developed by applying regularization to each effect using cross-validation. Regularization introduces a penalty term to reduce overfitting, with lambda values representing varying levels of regularization strength. Initially, the sequence of lambda values was set to seq(0, 10, 0.25), covering a broad range from no regularization (lambda = 0) to strong regularization (lambda = 10).

Each lambda value was tested to determine which one minimized the RMSE most effectively. This process allowed for the identification of the optimal level of regularization, striking a balance between reducing overfitting and preserving predictive accuracy. After determining that the optimal lambda was 5, the sequence was refined to seq(4.75, 5.25, 0.05) in subsequent runs. This adjustment reduced computation time by narrowing the focus to a more precise range of lambda values.

As shown in Figure 12, the optimal lambda value that minimized the RMSE was 5. The Regularized Model achieved an RMSE of 0.8626.

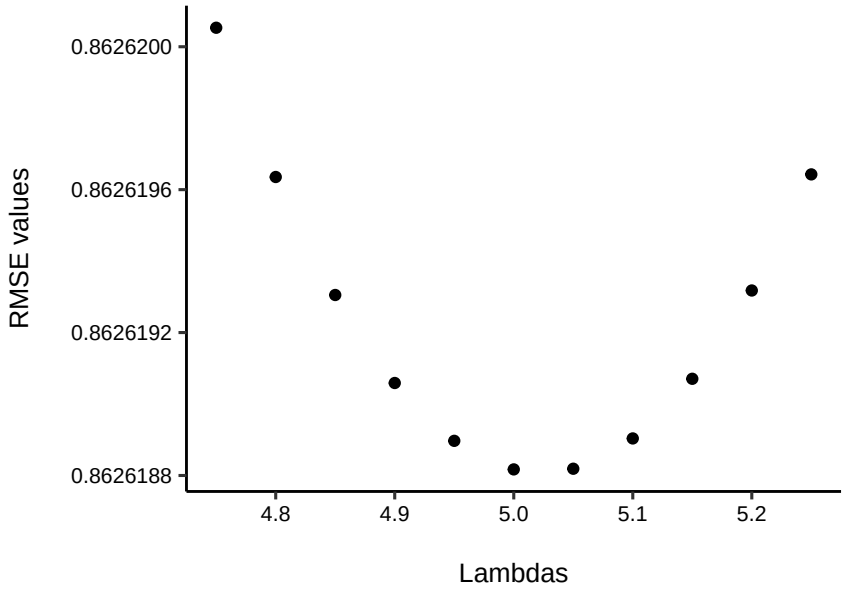

Figure 12: RMSE Against Lambda

Table 8 presents the RMSE values for all models developed, both with and without the clamping technique.

Table 8
RMSE Results including the Regularized Model

Model	Method	RMSE	Clamp
Model 1	Average Rating	1.0596	1.0596
Model 2	Movie Effect	0.9432	0.9432
Model 3	Movie + User Effects	0.8655	0.8653
Model 4a	Movie + User + Genre Effects	0.8652	0.8650
Model 4b	Movie + User + Genre-Specific Effects	0.8640	0.8638
Model 5	Movie + User + Genre-Specific + Release Year Effects	0.8637	0.8635
Model 6	Movie + User + Genre-Specific + Release Year + Review Date Effects	0.8635	0.8633
Model 7	Regularized Model	NA	0.8626

3.3 Final Model Assessment

In the final phase of the project, the algorithm was trained on the entire edx dataset, ensuring that all available data contributed to model development. This comprehensive training approach allowed the model to capture patterns more effectively by leveraging the complete set of movie ratings, user information, genre-specific features, release year, and review date effects.

After training the model on the edx dataset, it was applied to the independent evaluation dataset, final_holdout_test, to predict movie ratings. Before evaluation, the final_holdout_test dataset was modified to reflect the same transformations applied to the edx dataset, ensuring consistency in data structure between the training and evaluation phases.

The primary objective of this project was to develop a predictive algorithm that achieved an RMSE below the benchmark of 0.8649 on the final_holdout_test set. Meeting this threshold would demonstrate a high degree of accuracy in predicting movie ratings.

The final model produced an RMSE of **0.8623** for the unadjusted predictions. To enhance the model's realism, a clamping technique was applied, which restricted the predicted ratings to the valid range of 0.5 to 5. After clamping, the model achieved an RMSE of **0.8622** .

Figure 13 presents a histogram of the residuals (prediction errors), characterized by a bell-shaped curve. The distribution is centered around 0, indicating that the model's predictions are generally accurate, with most errors being close to zero. A significant portion of the predictions exhibit low errors, reinforcing the model's overall effectiveness in predicting movie ratings. The symmetrical decline in frequency as errors move away from zero further suggests that large deviations are relatively uncommon.

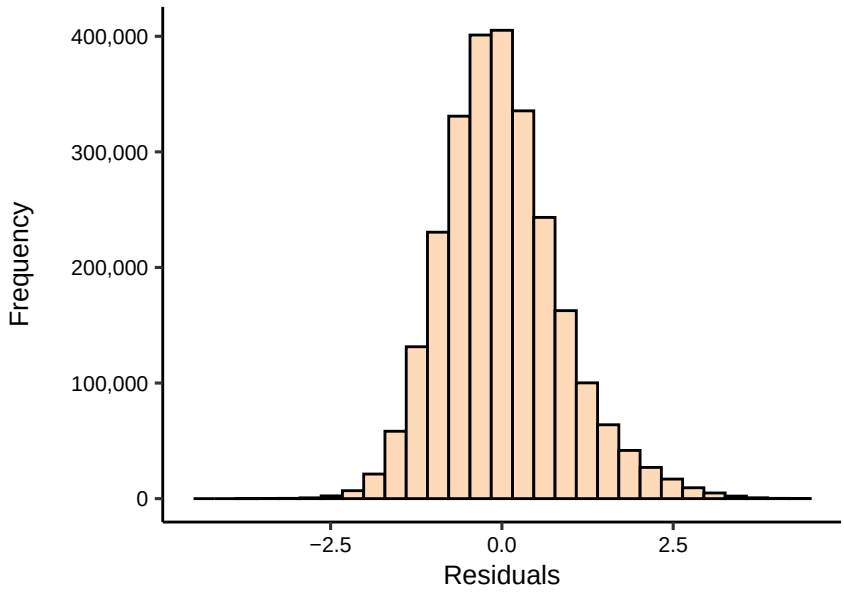

Figure 13: Histogram of Residuals

4 Discussion

The objective of this project was to develop an algorithm capable of accurately predicting movie ratings, with a target RMSE below 0.8649. The final model achieved an RMSE of **0.8622**, indicating that the model successfully met the desired accuracy threshold.

However, the techniques employed in this project were constrained by the computational limitations of training such a large dataset on a personal computer. While the models were able to perform well within these constraints, the accuracy and efficiency could potentially be improved with more advanced tools and computational resources.

Future improvements could involve leveraging more powerful algorithms and optimizing model parameters more effectively. For instance, matrix factorization techniques, such as Singular Value Decomposition (SVD) and Alternating Least Squares (ALS), are widely recognized for their ability to capture latent factors in rating data. Incorporating these methods could significantly enhance predictive accuracy and further reduce RMSE (Koren, Bell, & Volinsky, 2009).

5 Conclusion

In conclusion, the model successfully achieved an RMSE below the target threshold of 0.8649. While this outcome demonstrates the efficacy of the approach, further refinements could be made by incorporating advanced algorithms like matrix factorization to improve predictions. The MovieLens

project, though challenging, provided valuable insights into predictive modeling techniques and data handling on a large scale.

The extensive work invested in developing and refining the model enhanced both my technical and theoretical understanding, especially in managing large datasets and improving prediction accuracy. This project also deepened my practical experience gained through the Professional Certificate in Data Science, furthering my skills in data manipulation, modeling, and evaluation.

Additionally, generative artificial intelligence tools, including ChatGPT (OpenAI, 2024), were employed to support various aspects of the project. These tools facilitated the refinement of this written report, helped troubleshoot occasional coding issues, and streamlined workflow, improving the clarity of both analysis and presentation of results.

6 References

1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

3. Irizarry, R. A. (2019). Introduction to data science: Data analysis and prediction algorithms with R. CRC Press.

4. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37. https://doi.org/10.1109/MC.2009.263

5. Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). Applied Linear Statistical Models (5th ed.). Boston: McGraw-Hill Irwin, p. 248.

6. OpenAI. (2024). ChatGPT [Large language model]. https://chatgpt.com

7. Prat-Carrabin, A. & Woodford, M. (2022). Efficient coding of numbers explains decision bias and noise. Nat Hum Behav 6, 1142–1152, https://doi.org/10.1038/s41562-022-01352-4

8. R Core Team. (2023). R: A language and environment for statistical computing (Version 4.3.2). R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

9. Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. Springer. DOI: 10.1007/978-0-387-85820-3

10. RStudio Team. (2023). RStudio: Integrated Development for R (Version 2023.09.1). RStudio, PBC, Boston, MA. https://www.rstudio.com/

11. Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. Science, 185(4157), 1124-1131