# Developing and Evaluating a Movie Rating Prediction Model

Kevin Camp

2022-11-22

## 1.0 Introduction

This report satisfies the guidelines for the MovieLens project in the Data Science Capstone course. The project goal is analyzing a dataset of movie ratings by a collection of users in order to create a recommendation system. The recommendation system will predict ratings by user for movies which that user has not yet rated. The dataset is large; that presents a challenge, in that large datasets are difficult to manage. It also presents an opportunity, in that every data point can be used as a predictor for each missing value in the overall set. The aim of the project is to achieve that using machine learning methods.

## 1.1 Overview

The project uses a modified version of the MovieLens data with two sets. One set—"edx"—is used to train the model, and contains more than 9 million rows of user–movie ratings. Each row provides columns detailing a user identifier (userId); movie identifier (movieId); movie rating given by the user (rating); time in seconds since the Unix Epoch on midnight, January 1st, 1970 (timestamp); name of the rated movie (title); and genre or genres describing the rated movie (genres). The other set—"validation"—is similar in content to edx, and is used as the final hold-out test set. Table 1.1 presents a selection of rows from edx.

Table 1.1. First ten rows of edx dataset

| userId | movieId | rating | timestamp | title | genres |
|---|---|---|---|---|---|
| 1 | 122 | 5 | 838985046 | Boomerang (1992) | Comedy\|Romance |
| 1 | 185 | 5 | 838983525 | Net, The (1995) | Action\|Crime\|Thriller |
| 1 | 292 | 5 | 838983421 | Outbreak (1995) | Action\|Drama\|Sci-Fi\|Thriller |
| 1 | 316 | 5 | 838983392 | Stargate (1994) | Action\|Adventure\|Sci-Fi |
| 1 | 329 | 5 | 838983392 | Star Trek: Generations (1994) | Action\|Adventure\|Drama\|Sci-Fi |
| 1 | 355 | 5 | 838984474 | Flintstones, The (1994) | Children\|Comedy\|Fantasy |
| 1 | 356 | 5 | 838983653 | Forrest Gump (1994) | Comedy\|Drama\|Romance\|War |
| 1 | 362 | 5 | 838984885 | Jungle Book, The (1994) | Adventure\|Children\|Romance |
| 1 | 364 | 5 | 838983707 | Lion King, The (1994) | Adventure\|Animation\|Children\|Drama\|Musical |
| 1 | 370 | 5 | 838984596 | Naked Gun 33 1/3: The Final Insult (1994) | Action\|Comedy |

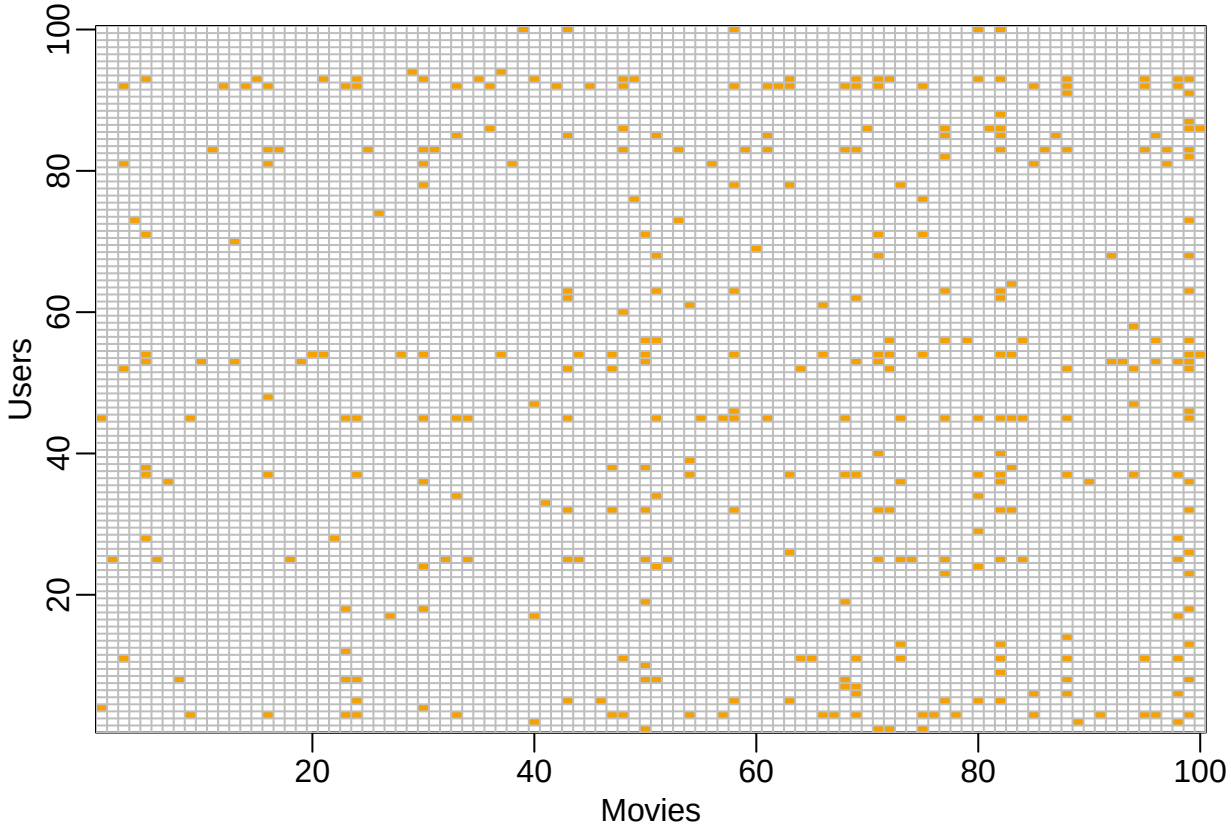
Additional summary detail on the edx set is included in table 1.2. The table shows the large number of users, movies, and movie genres represented in the data.

Table 1.2. Summary of edx set

| Rows | Columns | Unique users | Unique movies | Average rating | Number of genres |
|---|---|---|---|---|---|
| 9000055 | 6 | 69878 | 10677 | 3.51 | 797 |

Note the result from multiplying the unique users and unique movies values from table 1.2, 69,878 × 10,677 = 746,087,406, a number larger than the total row count for the edx set (9,000,055). This implies that in the edx data, each user did not rate each movie. It is then useful to view the edx dataset as a matrix, with users represented by rows and movies represented by columns. Figure 1.1 takes a random sample of 100 movies and 100 users and depicts user–movie combinations for which ratings exist with yellow shading. The machine learning algorithm developed in this project will allow us to use values in the entire matrix as predictors for each empty cell.

Figure 1.1. Matrix of a random sample of 100 users and 100 movies from edx




The edx data contains other useful pieces of information. For example, the timestamp column measures the timing of when users submitted ratings. The rating submission dates in the edx set range from 1995 to 2009, as illustrated in table 1.3.

Table 1.3. Date range in edx set

| Date of first rating | Date of last rating |
|---|---|
| 1995-01-09 | 2009-01-05 |

Another example is the genre information for each movie. Exploratory data analysis reveals at least one genre listed for the vast majority of movies recorded in the edx data. In fact, only one movie in the dataset—rated by 7 users—lacks a genre description as demonstrated in table 1.4. Referring back to table 1.1, we see that many movies have multiple genres listed. These variables, along with the others previously mentioned, will be useful predictors in the model.

Table 1.4. Rows in edx lacking genre information

| userId | movieId | rating | timestamp | title | genres |
|---|---|---|---|---|---|
| 7701 | 8606 | 5.0 | 1190806786 | Pull My Daisy (1958) | (no genres listed) |
| 10680 | 8606 | 4.5 | 1171170472 | Pull My Daisy (1958) | (no genres listed) |
| 29097 | 8606 | 2.0 | 1089648625 | Pull My Daisy (1958) | (no genres listed) |
| 46142 | 8606 | 3.5 | 1226518191 | Pull My Daisy (1958) | (no genres listed) |
| 57696 | 8606 | 4.5 | 1230588636 | Pull My Daisy (1958) | (no genres listed) |
| 64411 | 8606 | 3.5 | 1096732843 | Pull My Daisy (1958) | (no genres listed) |
| 67385 | 8606 | 2.5 | 1188277325 | Pull My Daisy (1958) | (no genres listed) |

## 1.2 Executive summary

Through data science techniques, we can create a strong prediction model capable of recommending movies to individuals. This can be beneficial for providing entertainment suggestions to individuals. An important question is: how can we model this efficiently with a large dataset? Machine learning methods are capable of achieving this, and that is the strategy I pursue in this report.

My project makes use of a large dataset. In the sections that follow I discuss my methods, clean and organize my data, develop my model, and finally evaluate the predictions generated by the model. I find success in predicting movie ratings based on four variables present in the dataset. The result is a suitably low measure of difference between the predictions my model generates and the actual rating values in a tranche of the dataset.

# 2.0 Methods and analysis

For this project, my goal is to develop a movie recommendation model. This model will aim to accurately predict movie ratings in the final hold-out test set as if they were not known to me. The means of evaluating my success will be minimizing residual mean squared error (RMSE). This project has a defined target RMSE value of less than 0.8649. As such, my goal is to design a model using machine learning methods that achieves RMSE $< 0.8649$.

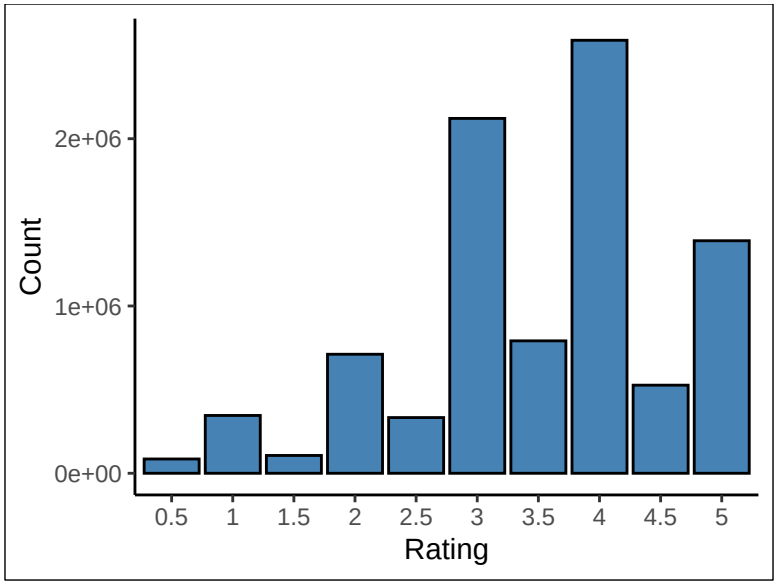
The key dependent variable of interest for recommending movies to users is the rating. The rating measures how much a user enjoyed a given movie. In the edx data, ratings range from a minimum of 0.5 to a maximum of 5.0. The overall mean value for the collection of user–movie ratings in the dataset is 3.51, and the standard deviation is 1.06. Table 2.1 displays these summary statistics.

Table 2.1. Summary statistics for ratings

| n | mean | sd | min | max | range | se |
|---|---|---|---|---|---|---|
| 9000055 | 3.512465 | 1.060331 | 0.5 | 5 | 4.5 | 0.0003534 |

Further exploring the data reveals additional patterns in user ratings. As shown in figure 2.1, we see that the most common rating is 4.0, and that whole integer values are more common in the data than values ending in a half star.

Figure 2.1. Count by rating in edx




Accurate prediction of movie ratings for a given user provides justification for recommending movies to that user. In brief, if my model predicts a user will assign a rating of 0.5 to a given movie, I would not recommend the user watch the corresponding movie. The reverse is true if my model predicts a user will assign a rating of 5.0.

## 2.1 Techniques and processes

My techniques and processes will include the following: 1. cleaning the data, 2. exploring and visualizing the data using the tidyverse package, 3. using insights gained to develop a model concept, 4. creating the model, and 5. evaluating the model on the validation set by measuring the RMSE.

### 2.1.1 Data cleaning

The timestamp in edx represents seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970. My first data cleaning effort was converting this variable to a date, which is easier to interpret and analyze. Table 2.2 displays the first five rows of the updated edx set with timestamp converted to a date format in the "rating_time" column and an added column, "rating_year" listing only the year of each movie–user rating.

Table 2.2. First five rows of updated edx dataset

| userId | movieId | rating | title | genres | rating_time | rating_year |
|---|---|---|---|---|---|---|
| 1 | 122 | 5 | Boomerang (1992) | Comedy\|Romance | 1996-08-02 | 1996 |
| 1 | 185 | 5 | Net, The (1995) | Action\|Crime\|Thriller | 1996-08-02 | 1996 |
| 1 | 292 | 5 | Outbreak (1995) | Action\|Drama\|Sci-Fi\|Thriller | 1996-08-02 | 1996 |
| 1 | 316 | 5 | Stargate (1994) | Action\|Adventure\|Sci-Fi | 1996-08-02 | 1996 |
| 1 | 329 | 5 | Star Trek: Generations (1994) | Action\|Adventure\|Drama\|Sci-Fi | 1996-08-02 | 1996 |

Movie release years are included in edx, but they are attached to the end of the text of cells in the title column. This information is more useful as a standalone column. As a result, my second data cleaning effort is to separate the movie release dates into a new "release_year" column, displayed in table 2.3.

Table 2.3. First five rows of edx dataset with release year

| userId | movieId | rating | title | genres | rating_time | rating_year | release_year |
|---|---|---|---|---|---|---|---|
| 1 | 122 | 5 | Boomerang (1992) | Comedy\|Romance | 1996-08-02 | 1996 | 1992 |
| 1 | 185 | 5 | Net, The (1995) | Action\|Crime\|Thriller | 1996-08-02 | 1996 | 1995 |
| 1 | 292 | 5 | Outbreak (1995) | Action\|Drama\|Sci-Fi\|Thriller | 1996-08-02 | 1996 | 1995 |
| 1 | 316 | 5 | Stargate (1994) | Action\|Adventure\|Sci-Fi | 1996-08-02 | 1996 | 1994 |
| 1 | 329 | 5 | Star Trek: Generations (1994) | Action\|Adventure\|Drama\|Sci-Fi | 1996-08-02 | 1996 | 1994 |

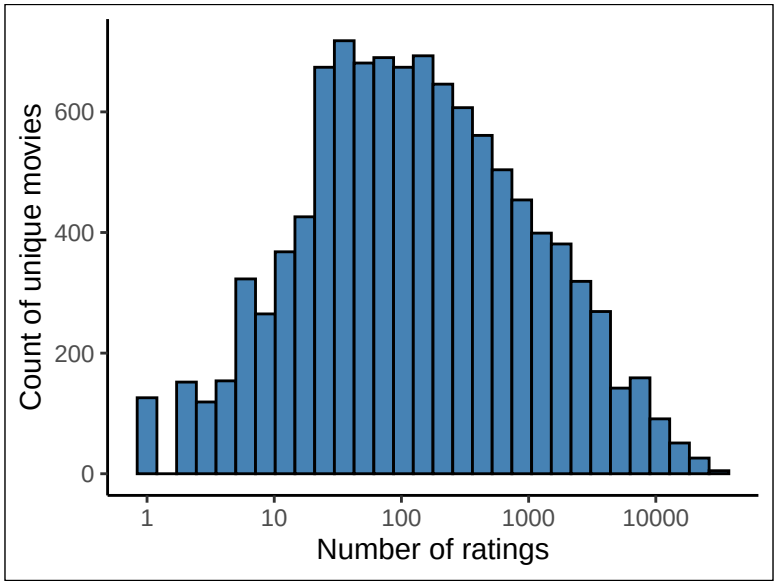
### 2.1.2 Data exploration and visualization

As demonstrated in table 1.2, the edx set contains 10,677 unique movies. Exploring these data, we find some of the movies receive more ratings than others. Summary statistics on ratings per movie are shown in table 2.4. Rating counts per movie range from 1 to 31362. Though the average ratings per movie is substantially high at nearly 843, this belies a notably large distribution; viewing this distribution of ratings per movie in histogram form in figure 2.2, we observe more than 100 movies in edx only received 1 rating, while a number of movies received more than 10,000 ratings.

Table 2.4. Summary statistics for ratings

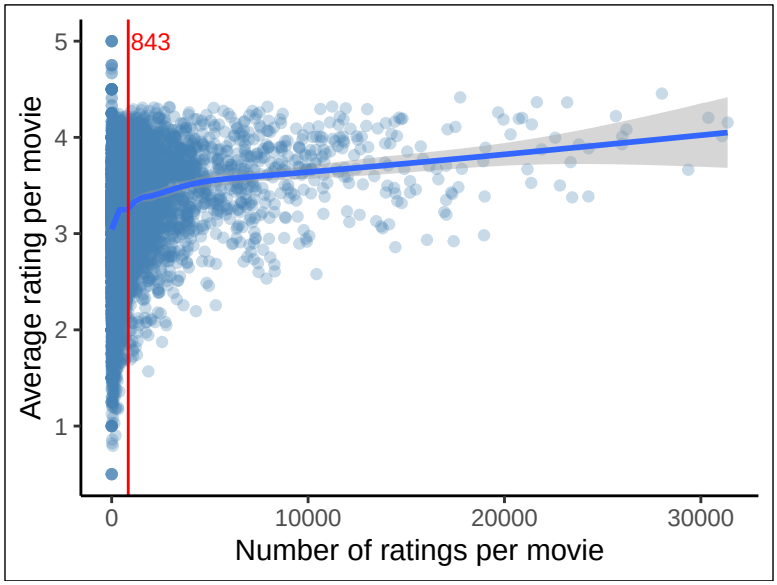
| n | mean | sd | min | max | range | se |
|---|---|---|---|---|---|---|
| 10677 | 842.9386 | 2238.481 | 1 | 31362 | 31361 | 21.66351 |

Figure 2.2. Distribution of ratings per movie



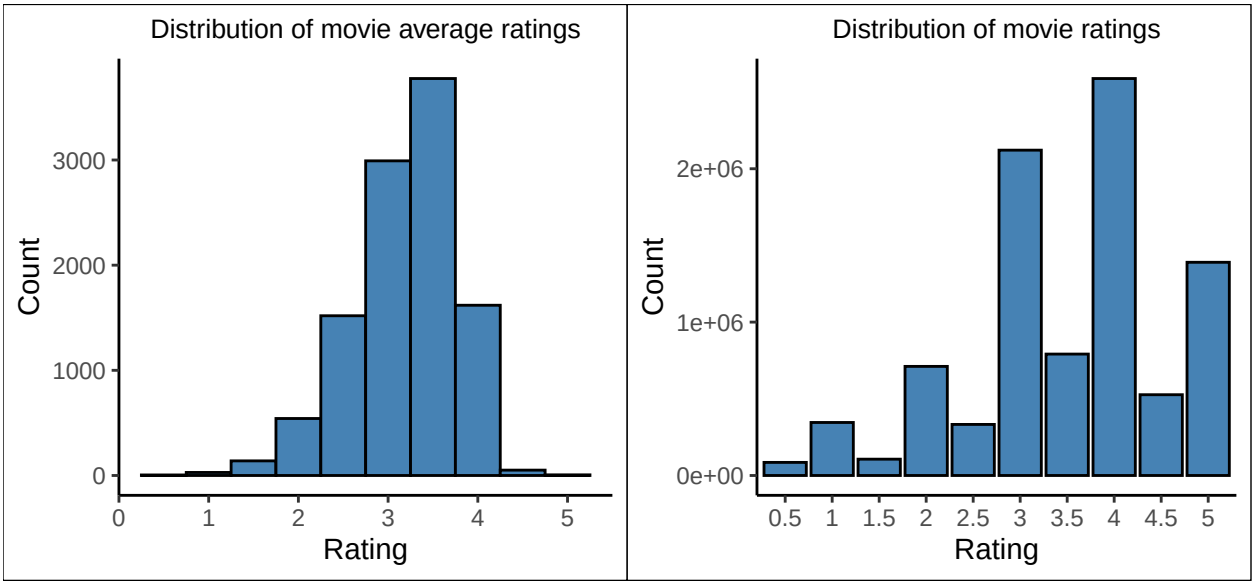

A large number of movies with few ratings presents a challenge for prediction accuracy. This is further illustrated in figure 2.3, which plots movies' average ratings by number of ratings. There, we observe that movies with fewer ratings vary widely—they may be rated close to 0.5 on average, or close to 5.0. By contrast, variability narrows for movies with more ratings—they tend to be rated closer to 4.0, on average. The trendline in figure 2.3 also indicates that average rating increases as the number of ratings increases. These observations will prove useful in creating the prediction model.

Figure 2.3. Average movie ratings by number of movie ratings




We can also investigate the distribution of average ratings by movie and compare it to the overall distribution of ratings. As we see in figure 2.4, the relative proportion of 5.0 ratings (seen in the right hand chart) greatly exceeds the number of movies that average a rating of 5.0 (seen in the left hand chart). This suggests user-to-user variability in rating behavior—for example, a given user could be especially inclined to give 5.0 ratings simply because they consider themselves a "positive person." On average however, this hypothetical effect would be abated.

Figure 2.4. Average and overall ratings distributions



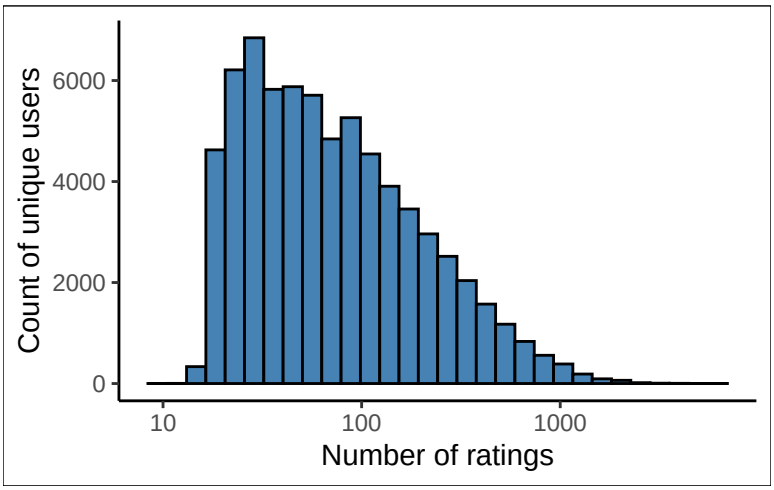

The 69,878 unique users in the edx data represent another variable worth examining. Exploratory analysis reveals that, just as ratings varied by movie, ratings vary by user as well. User rating summary statistics are presented in table 2.5. The lowest number of ratings submitted by a single user is 10, and the highest is 6,616.

The average user submitted nearly 129 ratings, but the large standard deviation of roughly 195 indicates substantial spread in ratings across individuals. The right-skewed distribution of ratings by user depicted in histogram form in figure 2.5 indicates many users submitted a low-to-moderate number of ratings, while a few users submitted very many.

Table 2.5. Summary statistics for users

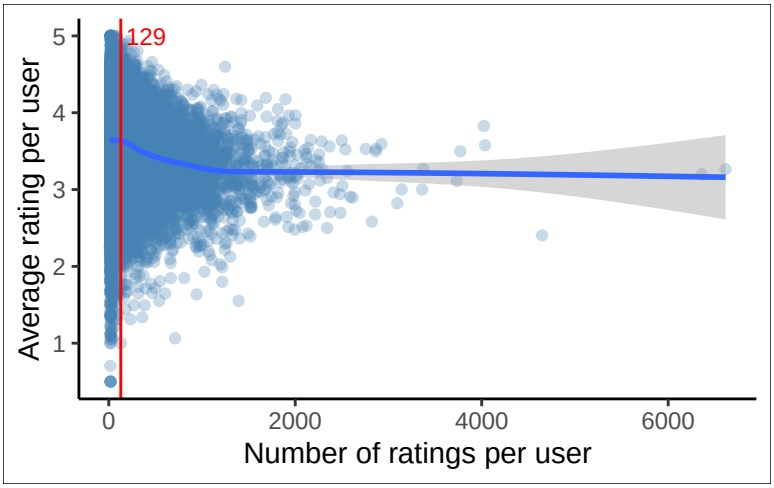
| n | mean | sd | min | max | range | se |
|---|---|---|---|---|---|---|
| 69878 | 128.7967 | 195.0602 | 10 | 6616 | 6606 | 0.7379015 |

Figure 2.5. Distribution of ratings per user



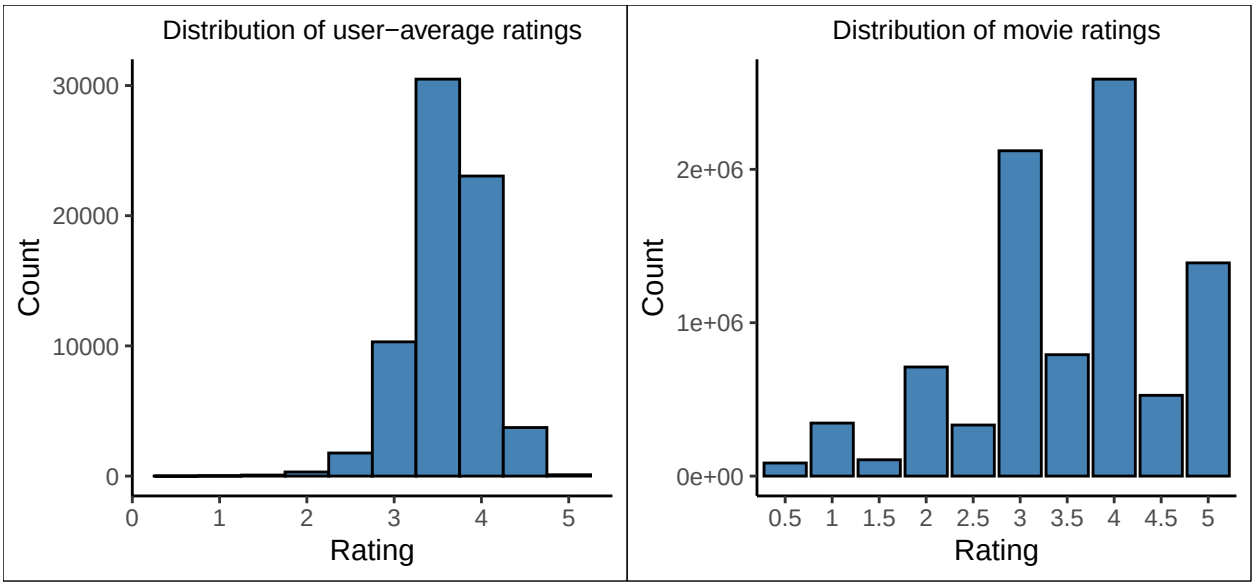

As with ratings per movie, variability in ratings per user complicates making accurate predictions. Figure 2.6 shows a scatterplot of average ratings by number of movies rated per user. The figure illustrates that less-active users vary greatly in their rating behavior, on average. Users with few total ratings may give movies nearly perfect scores on average, or conversely very low scores. This variability shrinks as ratings per user increase; more-active users appear to rate movies close to a 3.0 on average. A trendline reveals that average ratings decrease slightly as the number of user–ratings advances.

Figure 2.6. Average user ratings by number of user ratings



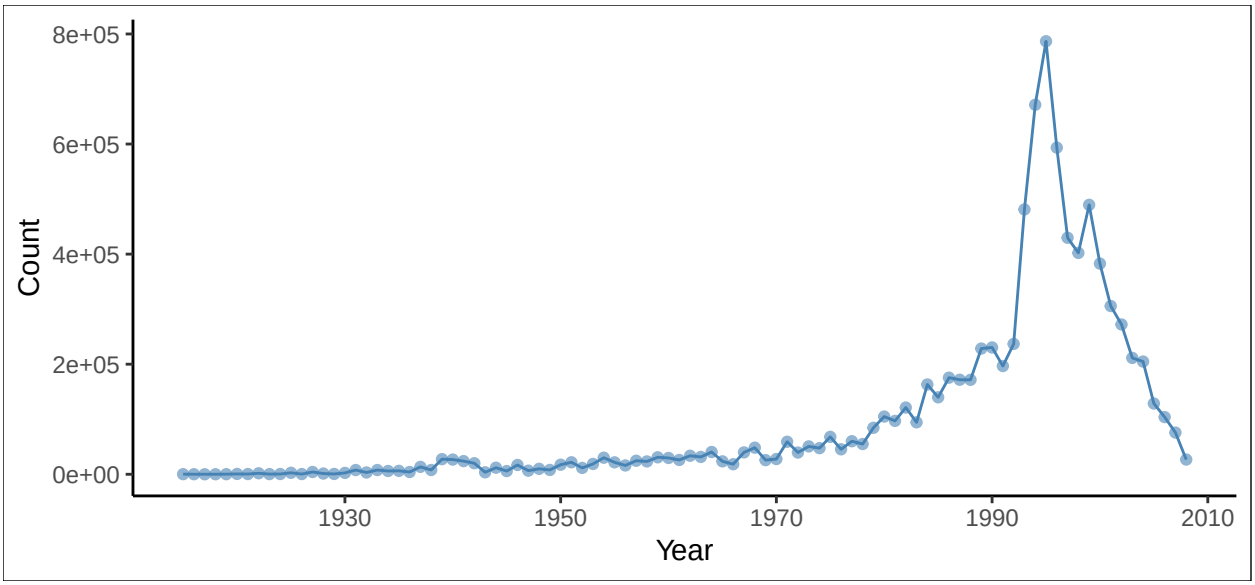
The distribution of average ratings by user differs from that of ratings overall. Figure 2.7 shows a modal value associated with lower ratings on the user–average (left hand) side. Another discrepancy is observed with 5.0 ratings—very few users average 5.0 for all movies they rated, yet the overall rating data contain many 5.0 ratings. These observations reinforce the expected conclusion that not all users—nor movies—are created equal.

Figure 2.7. User–average and overall ratings distributions



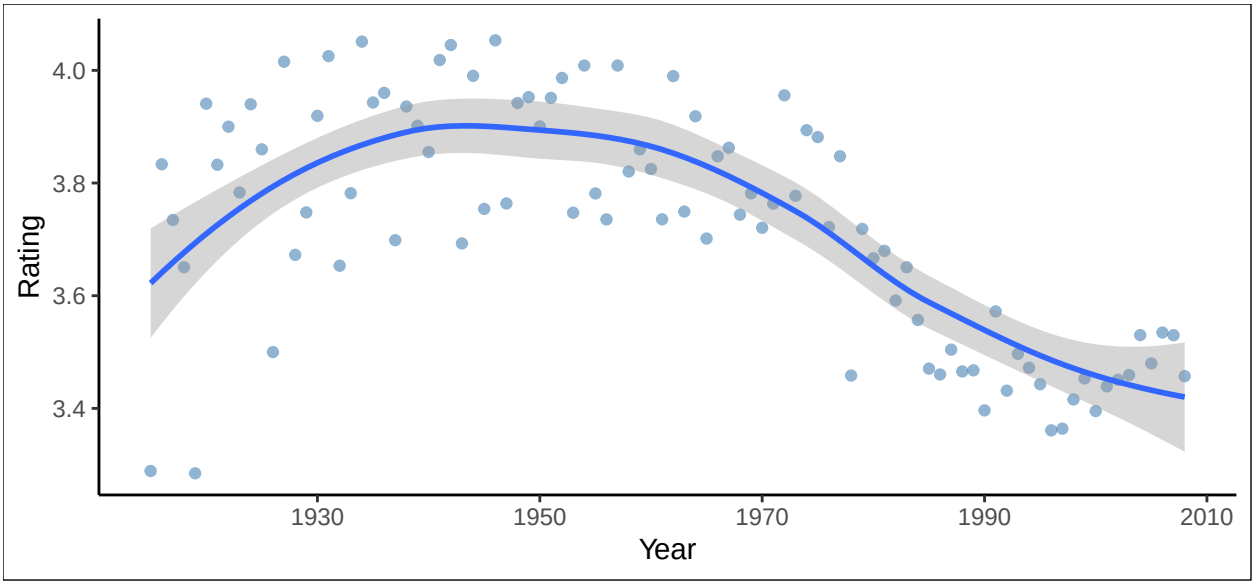

Do movie ratings in edx vary across release dates? Do rating patterns change in relation to the timing of submission? Exploratory analysis of the time trends in the data allows us to answer these questions. In figure 2.8, we see that ratings per year rise slowly at first, dating back to the early 1900s. Then yearly rating counts begin rising at an accelerated rate in the late 20th century, peaking in the mid 1990s. Since then, ratings per year decrease sharply.

Figure 2.8. Number of ratings by release year



Movies released in the mid 1990s may receive the most ratings, but are they rated the highest? Figure 2.9 suggests not; from this plot of average ratings by release year, we note that movies from the 1940s and 1950s appear to garner the highest acclaim. After the mid-century period, ratings declined consistently.

Figure 2.9. Average rating by release year




A simple linear regression model provides more information on the relationship between release year and average rating. Two models are estimated and results reported in table 2.6. Model 1 estimates the effect of the release year and squared release year on average rating; the squared term is included to account for the curved shape of average ratings by release year illustrated in figure 2.9. In model 2, a cubic term is added for comparison because the curve of average ratings by release year appears to have a point of inflection. The significant coefficients in both models indicate that release year impacts average rating, and that the impact varies over time.

Table 2.6. Linear models of average rating by movie release year

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | -580.88 *** | -27304.96 *** |
| | (83.25) | (6289.80) |
| I(release_year^2) | -0.00 *** | -0.02 *** |
| | (0.00) | (0.00) |
| I(release_year) | 0.60 *** | 41.48 *** |
| | (0.08) | (9.62) |
| I(release_year^3) | | 0.00 *** |
| | | (0.00) |
| N | 94 | 94 |
| R2 | 0.58 | 0.65 |

*** p < 0.001; ** p < 0.01; * p < 0.05.

In order to better utilize the time component of movie ratings, I add a new variable to the edx dataset to measure the time since a movie was first rated. This variable is computed as a simple subtraction of the earliest rating from a given user–movie rating. For ease of interpretation, I convert the resulting count of days into weeks via dividing by 7. Figure 2.10 shows the number of ratings since the first rating by week. Not surprisingly, rating numbers per movie peak near the time of that movie's first rating, then tend to taper off as time passes. This suggests initial excitement builds around a movie, which then abates as the movie ages and new films assume the spotlight.

Figure 2.10. Count of ratings, by week since first rating

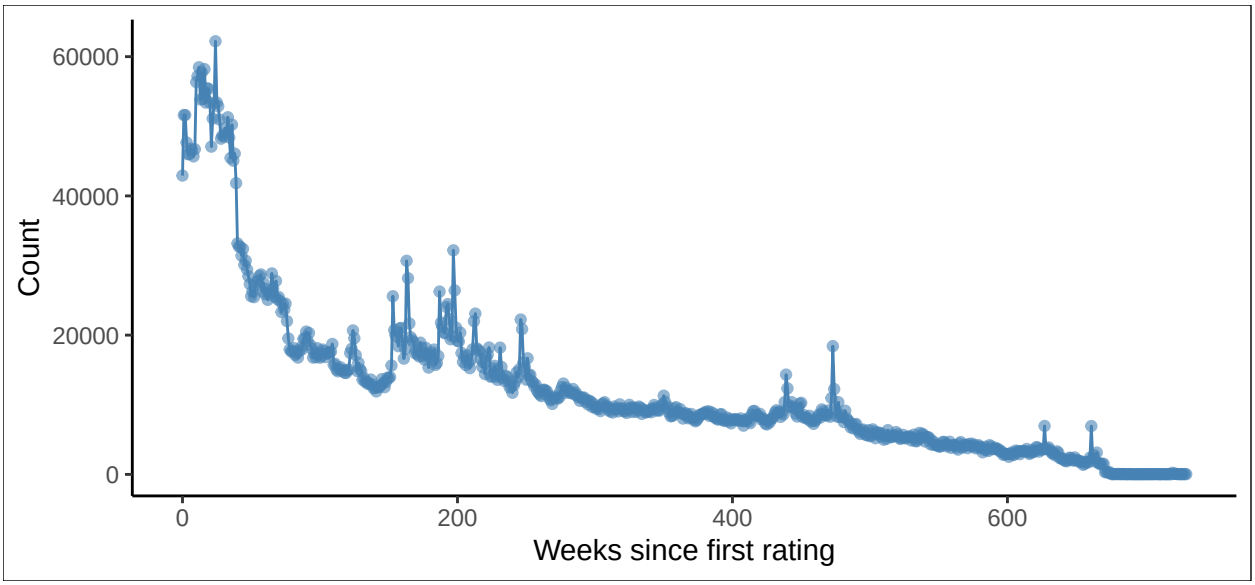
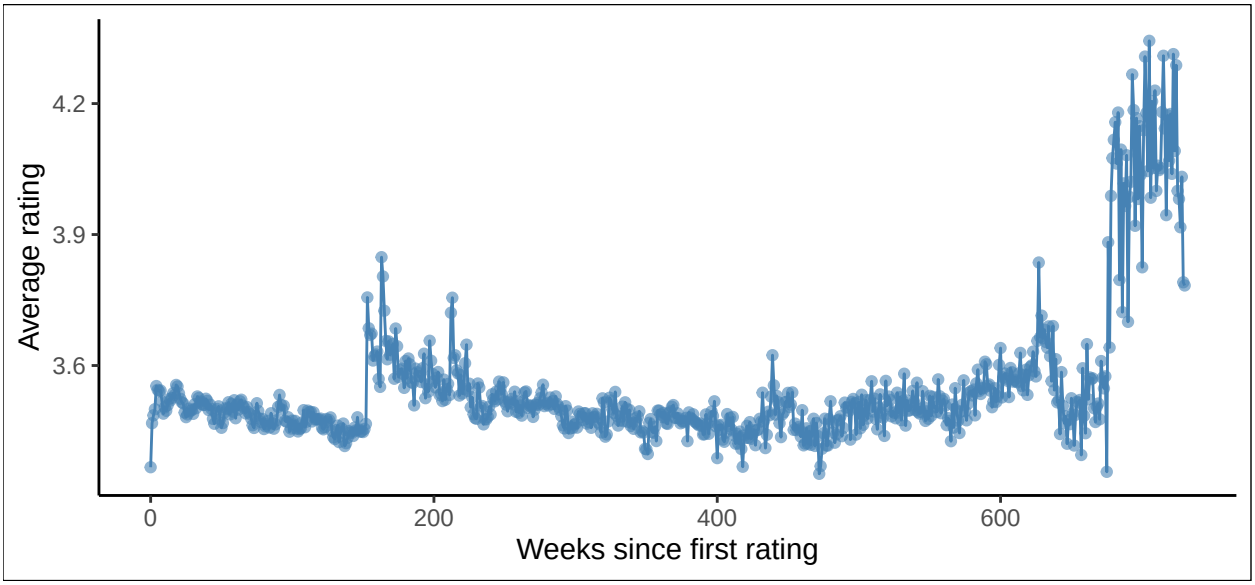


Figure 2.11 plots average ratings by time in weeks since a movie's first rating was submitted. The firgure shows that after an initial ratings bump (presumably during a period of buzz surrounding a newly released film), viewers tend to cool on movies and ratings decline slowly. As movies age since their first rating, average ratings step upward, perhaps due to nostalgia around the 3-year mark after release. Following that upward jump is another slow and steady average ratings decline. Years after the initial rating, average ratings increase substantially, but numbers of ratings are small; one could speculate that the cause is a subset of users who watch old films reverently.

Figure 2.11. Average rating, by week since first rating




Movies in the edx data are categorized into various genres. The distinct genres included in the dataset are listed in table 2.7. There are 19 unique genres (as well as the aforementioned "no genres listed" indicator). Notably, as demonstrated previously, a given movie can have more than one associated genre.

Table 2.7. List of unique genres in edx

| Genre |
|---|
| Comedy |
| Action |
| Children |
| Adventure |
| Animation |
| Drama |
| Crime |
| Sci-Fi |
| Horror |
| Thriller |
| Film-Noir |
| Mystery |
| Western |
| Documentary |
| Romance |
| Fantasy |
| Musical |
| War |
| IMAX |
| (no genres listed) |

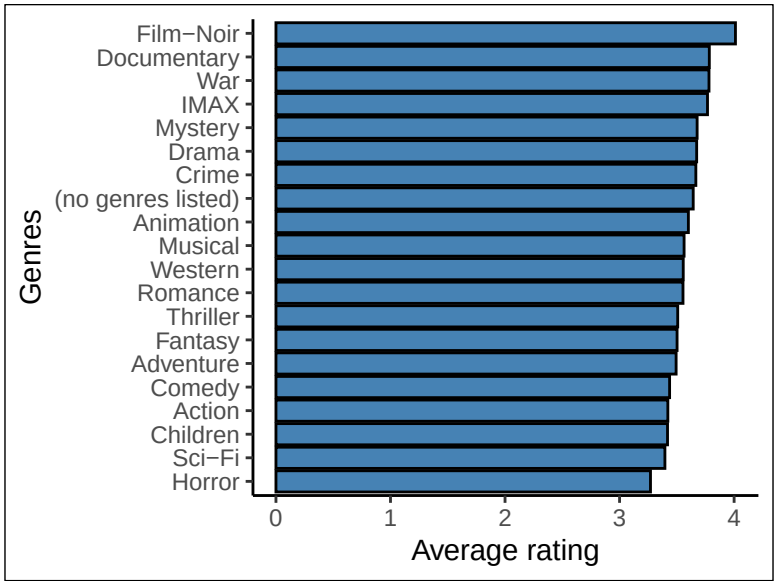
Table 2.8 displays the results of aggregating the most commonly mentioned genres in the edx data, listing in addition the average rating associated with movies of those genres. The most common film category is drama, representing nearly 4 million user–film observations in the set. The table demonstrates that ratings vary across genres, with drama films rating above action films by 0.25, on average.

Table 2.8. Most common genres in edx

| Genre | Movies | Average rating |
|---|---|---|
| Drama | 3910127 | 3.67 |
| Comedy | 3540930 | 3.44 |
| Action | 2560545 | 3.42 |
| Thriller | 2325899 | 3.51 |
| Adventure | 1908892 | 3.49 |

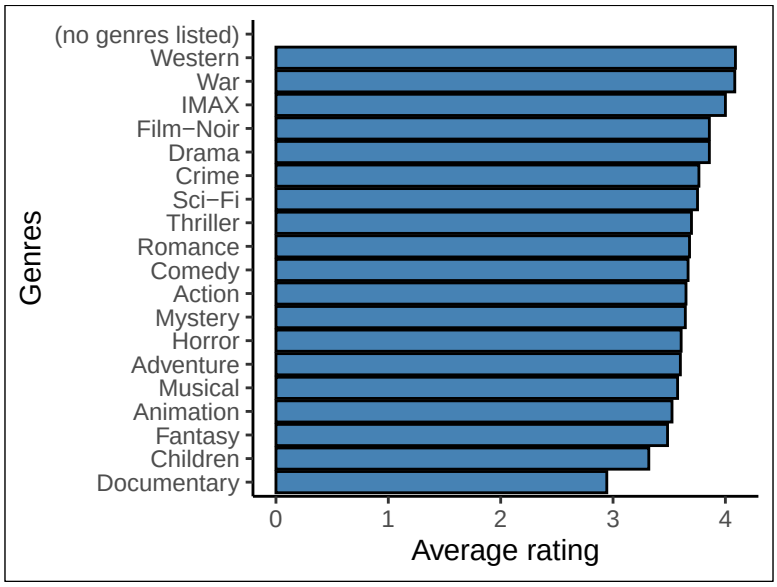
Film noir movies are the highest rated on average among the edx genres. Figure 2.12 shows they rate close to 4.0 on average. Conversely, the lowest rated films on average belong to the horror genre. The figure provides further evidence that ratings vary by genre.

Figure 2.12. Average rating by genre




Do genre effects interact with user effects? To investigate, we can take a random sample of 10 users from edx and explore how they rate movies on average across genres. In figure 2.13 we see the average ratings by genre for our random sample. The average ratings and the ranking of genres differs in figure 2.13 compared to figure 2.12. This suggests that, although there may be overall favorite genres for movie watchers, certain users have distinct genre preferences. These genre preferences will help inform our final model.

Figure 2.13. Average rating by genre, random sample of users




### 2.1.3 Insights gained

The exercise of data exploration and visualization resulted in several useful insights. Ratings data are unbalanced—a large number of movies have very few ratings. As the number of times a movie has been

rated rises, average ratings increase and rating variability declines. Despite users submitting a large number of overall 5.0 ratings, very few movies averaged a rating of 4.5 or higher.

Substantial user-to-user variation is present in the edx data. Certain users rate very few movies, while others submitted thousands of ratings. User rating averages varied widely among those with few ratings. This variability was not observed for users with many ratings. In addition, very high and very low user–rating averages were rare, especially in light of how many 5.0 ratings were submitted overall.

The ratings of movies are correlated with their release years. After a movie receives its first rating, other ratings pour in—but not for long, and soon rating numbers decrease. Average ratings step upward, then steadily decrease with age, until years later when movies experience an apparent critical resurgence.

Finally, genres matter for movie ratings overall. More movies from certain genres—drama, for example—were reviewed in our dataset. Based on average rating, users preferred some genres over others. These genre effects seem to vary on a user-to-user basis.

## 2.2 Modeling approach

My approach to developing the model is to begin with a naive attempt using a simple average. From there, I will tune my model, adding predictors incrementally in an effort to minimize RMSE. I plan to use predictors based on the insights detailed above, to ultimately produce a model that accounts for movie, user, time, and genre effects on ratings.

Values of the key dependent variable, ratings, in edx are discrete. In other words ratings values are measured by counting. One improvement in my model is that it will allow for ratings predictions that are continuous. The result should be a more sensitive measure and, by extension, a lower RMSE.

# 3.0 Results

In this section I develop my model to minimize the RMSE of predicted movie ratings and actual ratings in the edx dataset, then present the results of my model. As previously described, I adopt an iterative approach, adding predictors to the model one-by-one.

## 3.1 Model results

The most straightforward model for a movie recommendation system would simply predict the same rating for each movie, not taking into account unique characteristics of the movie, the user rating it, or the time in which the rating was submitted. Such a model would use the average to minimize the RMSE. With the edx data, a model using only a simple average of movie ratings results in a RMSE of 1.06, as shown in table 3.1.

Table 3.1. RMSE by method I

| method | RMSE |
|---|---|
| Only estimate is the average | 1.0603 |

The RMSE of 1.06 is quite high—it would translate to regularly missing ratings predictions by more than a full point. We can refer back to table 2.1 to confirm that our current model RMSE is equal to the standard deviation of movie ratings in edx. A reasonable goal is to improve upon this RMSE. In an effort to do so, we can tweak the model so that instead of using a simple average as the estimate, we use a movie–average, *i.e.*, the average rating per movie. This will help disentangle the movie-specific effects we observed in the data

(some movies are "better" than others). Table 3.2 adds a new row to report the RMSE for our movie–average model.

Table 3.2. RMSE by method II

| method | RMSE |
|---|---|
| Only estimate is the average | 1.0603 |
| Add movie–average effect | 0.9423 |

We see that our updated model has improved RMSE, lowering it to 0.94. The next step is to account for user-specific effects. To compute the user-specific effect for a given movie, we subtract the movie–average rating from each user–movie rating (*i.e.*, row in edx). This provides a measure of how each user differed from the average rating for each movie. Table 3.3 adds another new row to show the results of the model incorporating an estimate for the user effect.

Table 3.3. RMSE by method III

| method | RMSE |
|---|---|
| Only estimate is the average | 1.0603 |
| Add movie–average effect | 0.9423 |
| Add user–specific effect | 0.8567 |

Incorporating an estimate for the effect of users improves our model performance once again. Now we need to address the variation of ratings over time in our data. We will add a time effect using the variable of weeks since a movie's first rating, the result of our data cleaning and exploratory data analysis. The resulting model RMSE is provided in table 3.4.

Table 3.4. RMSE by method IV

| method | RMSE |
|---|---|
| Only estimate is the average | 1.0603 |
| Add movie–average effect | 0.9423 |
| Add user–specific effect | 0.8567 |
| Add time effect | 0.8562 |

The measure of weeks since first rating represents an improvement to the model, decreasing RMSE from 0.8567 to 0.8562. The logical final step to round out the model is adding a treatment for genres. The result of incorporating genre effects into the model is depicted in Table 3.5.

Table 3.5. RMSE by method V

| method | RMSE |
|---|---|
| Only estimate is the average | 1.0603 |
| Add movie–average effect | 0.9423 |
| Add user–specific effect | 0.8567 |
| Add time effect | 0.8562 |
| Add genre effect | 0.8559 |

With all variables included, our model results in an RMSE of 0.8559 on the edx ratings data. The culmination of the project is to evaluate the model's performance on the final hold-out test set.

## 3.2 Model performance

The first step of evaluating the final model on the validation dataset is to repeat our data cleaning procedure in order to get the measure of weeks since a movie's first rating. Then, we will append columns for the four final model effects—movie, user, time, and genre. Finally, we will calculate RMSE, measuring success by whether the value is below 0.8649.

Before performing the final evaluation, we need to address not applicable (NA) values in the time variable predictor appended to the validation set. Table 3.6 reveals that the user and genre effect columns do not contain NAs, but the column containing measures of time in weeks since a movie's first rating does. To deal with these, I convert each NA to the average value of the time effect in the validation set.

Table 3.6. Check for NA values in validation set

| | NAs |
|---|---|
| b_u | 0 |
| b_t | 434 |
| b_g | 0 |

With the NA values converted, we can evaluate the final model. Results are shown in table 3.7.

Table 3.7. Final RMSE evaluation

| method | RMSE |
|---|---|
| Movie, user, time, genre effect model | 0.8645 |

My model's final RMSE is 0.8645. The effects included in the model allow us to make predictions that are generally within the target of 0.8649, meaning we can generally make predictions of movie ratings that are suitably close to the genuine values observed in the data.

# 4.0 Conclusion

This project involved making predictions on movie ratings using machine learning techniques. The contribution of this work is demonstrating how simple methods can result in strong predictive power on variables in large datasets. This is useful for individuals seeking entertainment—movies, for example. But there are many other potential practical applications for the techniques used in this analysis. I will discuss those after summarizing the report and addressing the limitations of this work.

## 4.1 Report summary

With the aim of developing an accurate movie recommendation model, I performed analysis on a large dataset. The data were millions of movie ratings with corresponding information on users, time, and genre. My efforts involved summarizing the data, cleaning the data, performing exploratory analysis and data visualization. The result was added insight which, in turn, I put to use while considering strategies for modeling movie ratings. I developed my model, electing for an iterative process to drive RMSE lower in increments. Adding in effects by movie, user, time, and genre allowed my model to account for the rich variability observed across all. Ultimately, I was able to achieve the goal of an appropriately low RMSE when evaluating my model's predictions compared to the ratings in the final hold-out test set.

## 4.2 Limitations

My project, while successful, is not without drawbacks. Movies with low rating counts confound predictions that use average rating as a benchmark, such as mine. A similar case can be made for users with low rating counts. Aside from that, my analysis uncovered some odd patterns in the data—for example, average movie ratings increasing substantially once roughly 600 weeks have passed since the first rating. This is an example of a phenomenon that is difficult to explain, and perhaps inadequately addressed in my analysis. Another possible limitation of this work involves the handling of NA values in the validation set. I opted to convert those to the column average for simplicity—it was straightforward. But it is possible that closer investigation would uncover a better way to deal with the NAs.

## 4.3 Future work

Future analysis would benefit from improved data. More granular genre categories could be one improvement, as there are likely more than 19 unique genres that are informative with respect to movies. Other improvements to the dataset could come by incentivizing users to rate more movies; the large number of users who rated comparatively few movies could be providing much more predictive power. Finally, it is likely that new or more advanced modeling techniques could push the RMSE down further. I met the target goal, but there is likely room for improvement in future studies.

It is easy to imagine applications of this work to areas outside of movie recommendations. Many everyday activities—shopping, food away from home, hotels, etc.—operate on rating systems. A natural extension of the work in this report would be predicting values users place on those experiences. This could result in better understanding of how people spend their time, which has implications not only for lifestyles but also policy. Future work would do well to explore this.