Capstone Project1 : MovieLens

Emmanuel Antiri

2022-06-25

1. Introduction

Feedback is an important component in building and sustaining organizations for the foreseeable future. For instance, valuable feedback could assist businesses in improving their processes, products, and strategies. In recent times, movie ratings have taken prominence in literature and in practice, inspired by the Netflix challenge. In this report, we build a recommendation system that predicts movie ratings using data obtained from the MovieLens dataset. The final model evaluated on the validation set achieved an RMSE of **0.864528** (lower than the threshold of 0.86490).

1.1 Description of Dataset

The MovieLens dataset was used to build the recommendation system. Using random sampling, the dataset was first split into validation set (10% of the MovieLens dataset) and edx set (the remaining 90% of the MovieLens dataset). The edx set was further split into train and test sets. The train set was 90% of the edx dataset whiles the test set was 10%. It should be noted that after each of the two splits, reasonable joins are done to ensure that same users and movies are present in both sets. The dataset is made up of six (6) variables: userId, movieId, rating, timestamp, title and genres. UserId and MovieId variables specify unique identification of users and movies respectively. The rating variable specifies the rating score given to a movie by a user. Timestamp specifies the specific time that the rating was given by the user. The title variable specifies the title and release year of the movie. Genre indicates the type of content of the movie.

1.2 Goal

The main goal of the project is to propose a recommendation system that achieves an RMSE below 0.86490. In doing so, the paper also demonstrate that the proposed algorithm is easily interpretative and computationally inexpensive. Moreover, the paper would showcase how individual, movie and time effects influence movie ratings.

1.3 Key Steps

The steps used in building the final recommendation system is outlined below: 1. The overall mean of ratings is used as base prediction and rating variations about the mean is recognized as the error term. 2. Movie effects is incorporated into the prior step. 3. Individual user effects is incorporated into the prior step. 4. Ascertaining whether regularization produces superior results. 5. Incorporating time effects (year of movie release and year of movie rating) into Step 3 because regularization did not produce much difference. 6. Incorporating the interaction term of year of movie release and year of movie rating into the prior step.

It should be noted at each step of the process, the performance of the model in terms of RMSE was evaluated using the test set. Finally, the final algorithm was evaluated on the validation set which achieved an RMSE lower than the threshold. The R software was used in performing the analysis.

2. Methods/Analysis

2.1 Data Cleaning

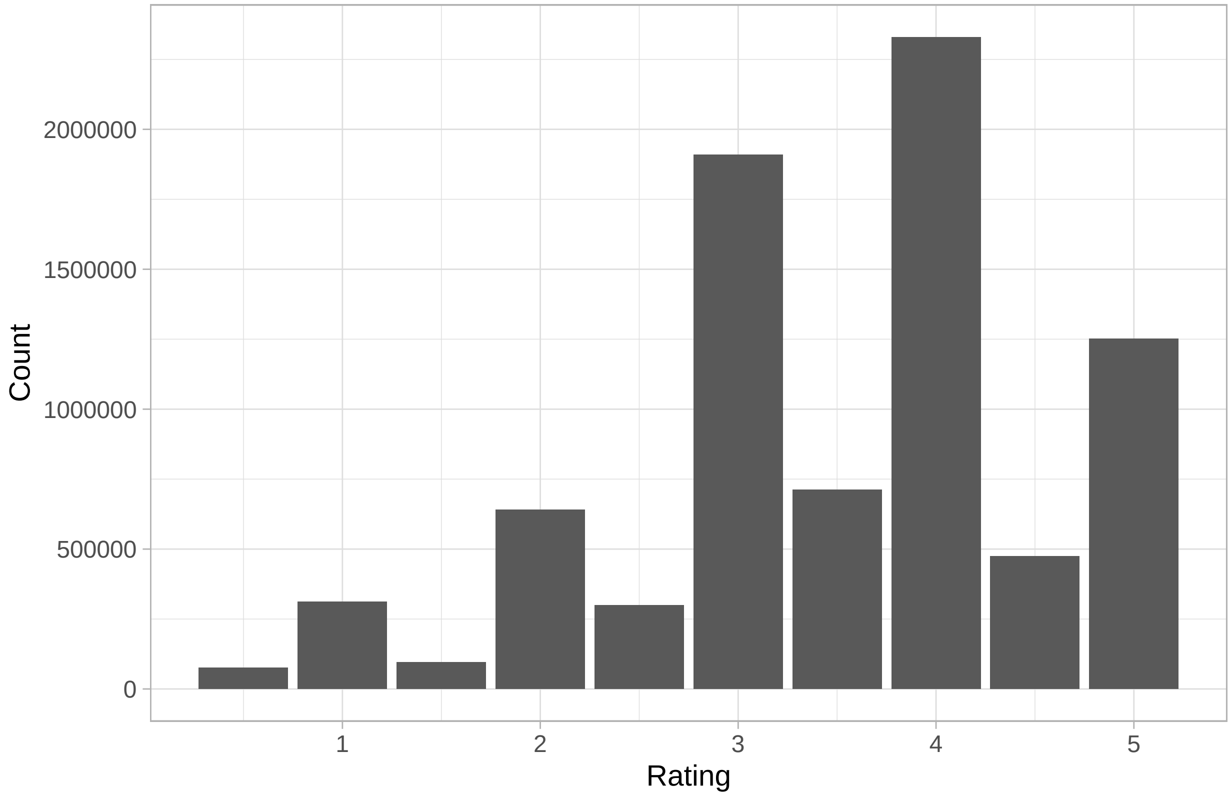
The train dataset was used to build the models that were evaluated on the test data. During splitting of data, reasonable joins were done to ensure that same users and movies are present in both sets. The dataset is made up of six (6) variables: userId, movieId, rating, timestamp, title and genres. Derived variables used in estimation included year of movie release, year of movie rating and interaction term of year of movie release and year of movie rating.

2.2 Data Exploration and Visualization

2.2.1 Movie Ratings

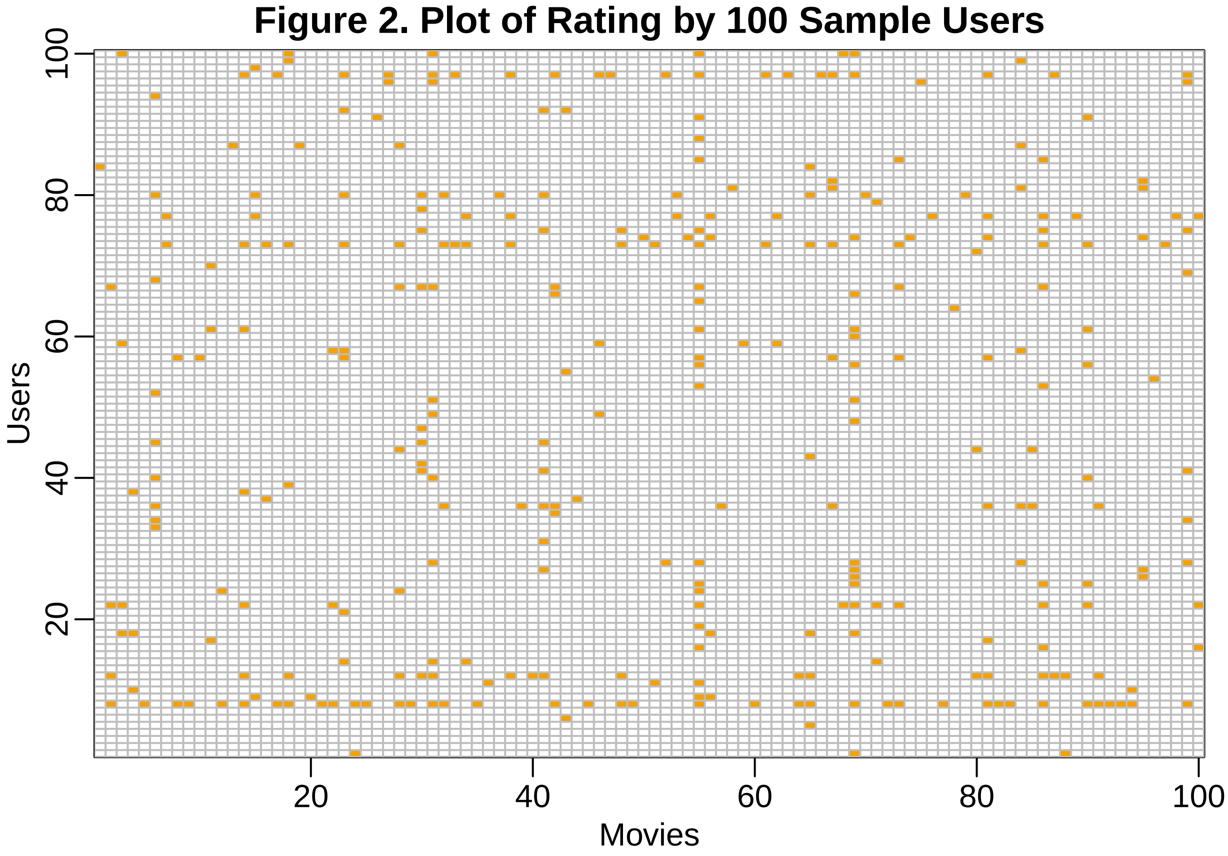
From the train set, we realize that 69878 unique users rated 10677 unique movies, with rating ranging from 0.5 to 5. Figure 1 below reviews some characteristics of the distribution of ratings. First, Figure 1 shows that users are more likely to give favourable ratings; the average rating is higher than the mid-point rating of 2.75 stars. The modal rating was 4 stars. Additionally, the lowest star rating of 0.5 had the least number of user ratings.

Figure 1. Distribution of Rating



2.2.2 Movie Ratings of Users

From the train set, we note that the number of users does not equal the number of movie reviews. Therefore, not all users watched/rated all movies, and possibly not all movies were rated by the same users. Figure 2 below shows evidence that not all movies were rated by the same users and not all users rated every movie. The few coloured indications on the plot show that users rated unequal number of movies.



2.2.3 Distribution of Users and Movie Ratings

The dissimilar pairing of the frequency of users as against number of movies rated is also represented in Figure 3 and Figure 4 below. In both plots, the number of similar group of users or movies is logarithmic transformed to base 10.

Figure 3 below represents the distribution of movies from the train set, showing an approximately mesokurtic distribution. The plot shows that most of the rated movies are within the middle-tier counts of movies, though there exist movies that are very popular or very rare. Moreover, Figure 3 shows that there are more rare movies than there are popular movies, though there are intervals within the lower band where no movies were rated.

Figure 3. Distribution of Movies

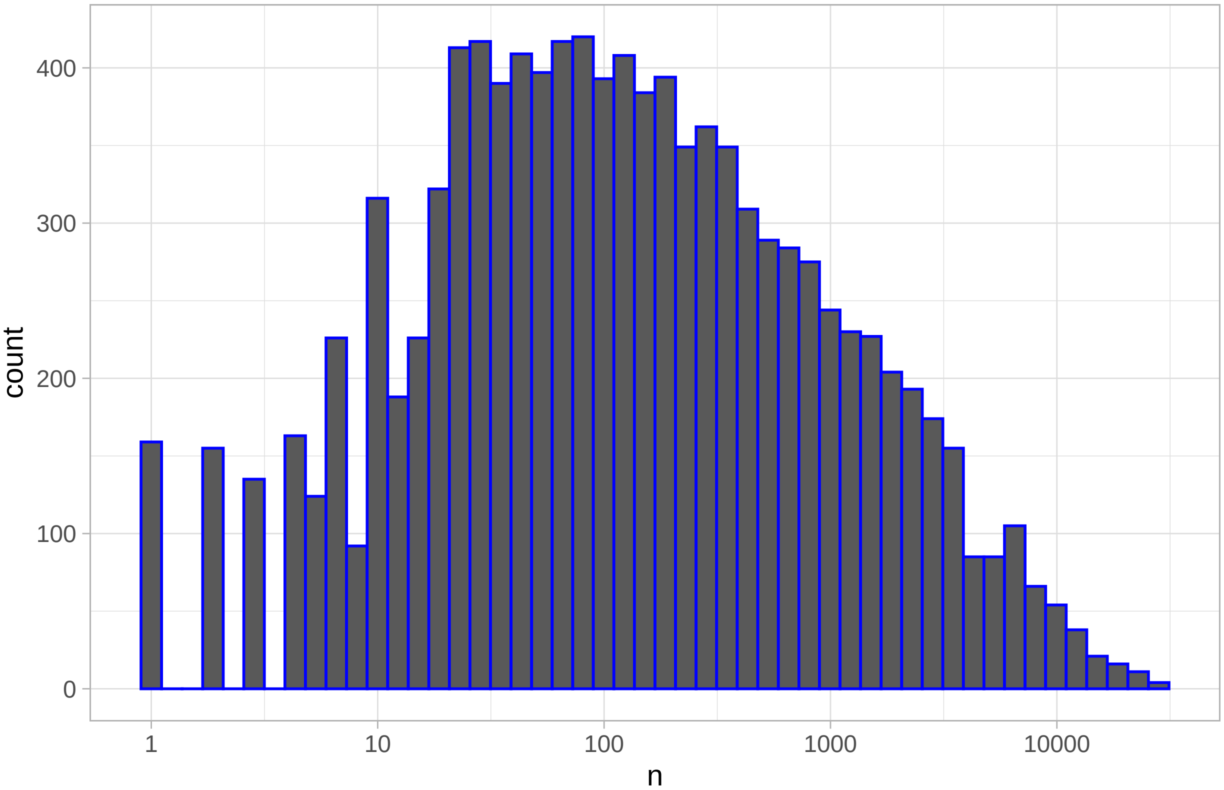
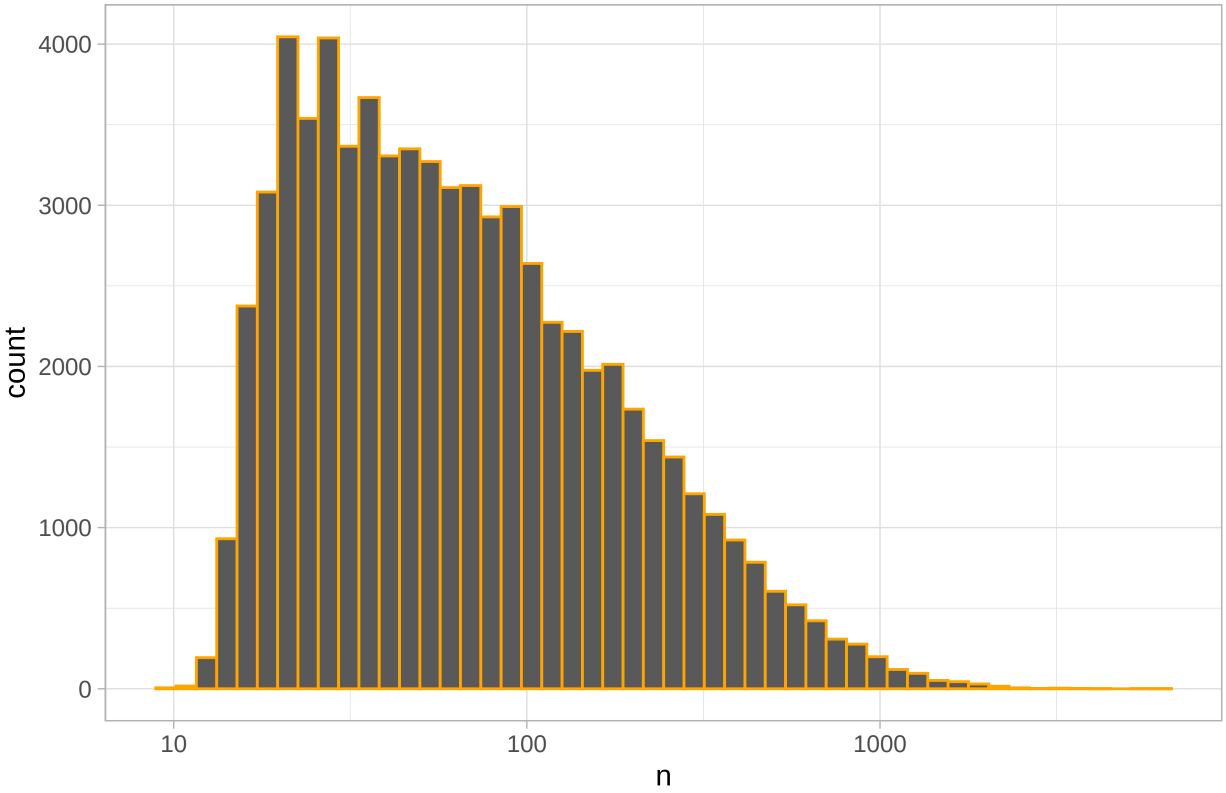


Figure 4 below represents the distribution of users from the train set, showing right-skewed distribution. The plot shows that most users rated few movies whiles few users rated many movies. In other words, the modal and average distributed value of users' engagement with movie rating is below the median value. The skewness of the distribution means that the median value will provide a better center of dispersion than the average value.

Overall, Figure 3 and Figure 4 provide some cues in subsequently modelling the prediction of ratings. First, the plots show that there exist different characteristics of users and movies and these effects ought to be examined further and incorporated into models that tend to predict movie ratings. Also, the evidence that tends to show that there exist unequal number of users and movie rating pairing are shown by combining both plots. Obviously, these distributions are more likely to exist in real life. It is very intuitive to realize that movie watching, and to a lesser degree, movie rating is skewed because not all movies become popular. In the next section, we would further examine the proposition that not all movies are popular by understanding the distributed rating of movies by genres.

Figure 4. Distribution of Users



2.2.4 Movie Ratings by Genre

In basic terms, the genre of a movie reflects the type of content to expect of the movie. Each movie can be categorized in at least one movie genre. There were 797 different combination of genres for movies contained in the train set. Table 1 and Table 2 provide the summary of the top ten (10) and the last ten (10) movies arranged by average ratings. The tables also reveal different frequency of ratings per movie. Thus, this difference could connote that movie with few reviews could have upward or downward biases in rating. The effects of the differing number of ratings would be examined in the modelling phase.

```
##                                    genres Frequency Average_rating
## 1                    Animation|IMAX|Sci-Fi         6         4.6667
## 2                  Drama|Film-Noir|Romance      2693         4.3036
## 3                  Action|Crime|Drama|IMAX      2095         4.2995
## 4          Animation|Children|Comedy|Crime      6418         4.2789
## 5                        Film-Noir|Mystery      5431         4.2393
## 6                  Crime|Film-Noir|Mystery      3650         4.2240
## 7               Film-Noir|Romance|Thriller      2190         4.2174
## 8                 Crime|Film-Noir|Thriller      4365         4.2115
## 9                   Crime|Mystery|Thriller     24173         4.2011
## 10 Action|Adventure|Comedy|Fantasy|Romance     13329         4.1971
```

```
##                                         genres Frequency Average_rating
## 788                   Action|Adventure|Children       745         1.9342
## 789                      Action|Children|Comedy       460         1.9315
## 790                 Action|Horror|Mystery|Sci-Fi        19         1.9211
## 791       Action|Adventure|Drama|Fantasy|Sci-Fi        51         1.9020
## 792 Adventure|Animation|Children|Fantasy|Sci-Fi       627         1.9003
## 793      Adventure|Drama|Horror|Sci-Fi|Thriller       196         1.7959
## 794                  Action|Drama|Horror|Sci-Fi         4         1.7500
## 795                   Comedy|Film-Noir|Thriller        17         1.6471
## 796              Action|Horror|Mystery|Thriller       289         1.6142
## 797                          Documentary|Horror       547         1.4415
```

2.2.5 Movie Ratings: Time Effects

In this section, we examine the time effects on ratings. Ratings has not been the same across year of rating (Figure 5) and year of movie release(Figure 6). In all cases, the average rating was above 3.20, an indication that individuals then to rate favorably than not. However, we note marked differences in trend based on movie released. First, variations in ratings of old movies tend to be higher than the current movies. The plot shown an inverted shape trend of movie released. Movies released between 1927 and 1978 were rated consistently above 3.65 on average. However, from the year 1979, the average rating of movies shown a downward trajectory.

Figure 5. Evolution of Average Ratings by Year Rated

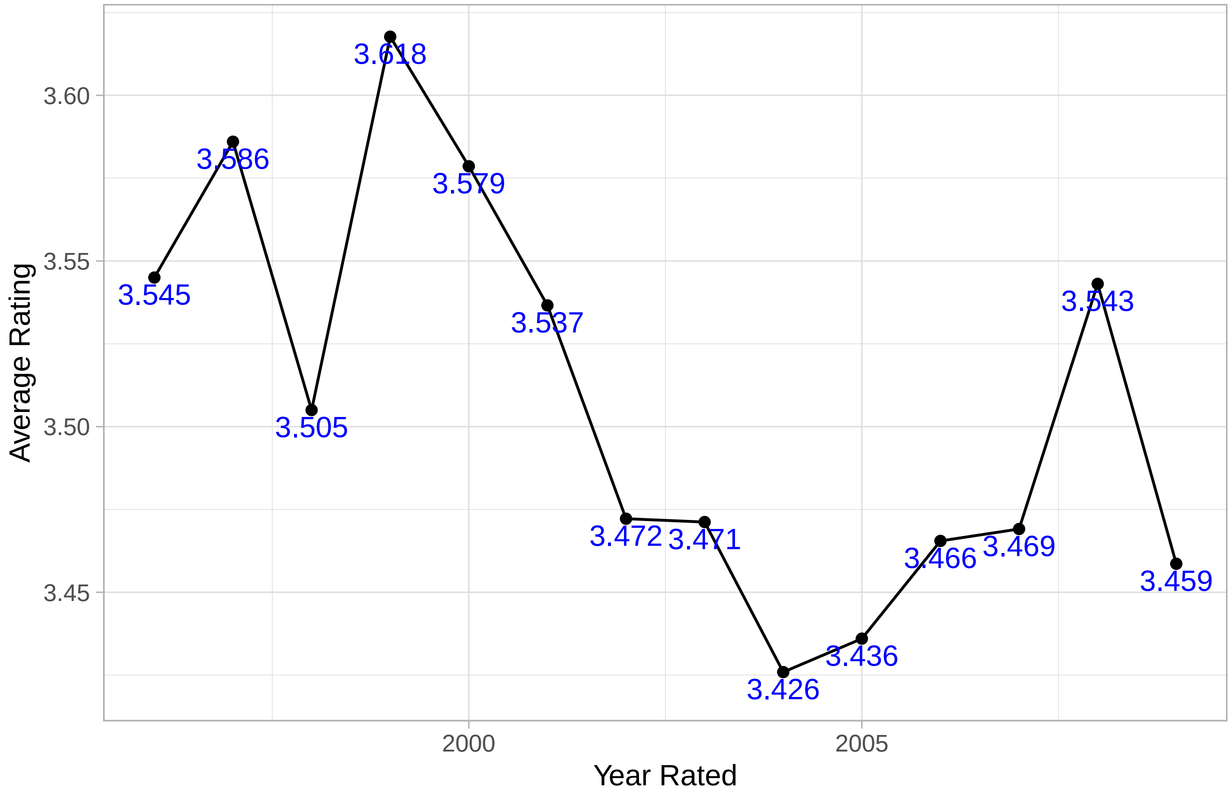
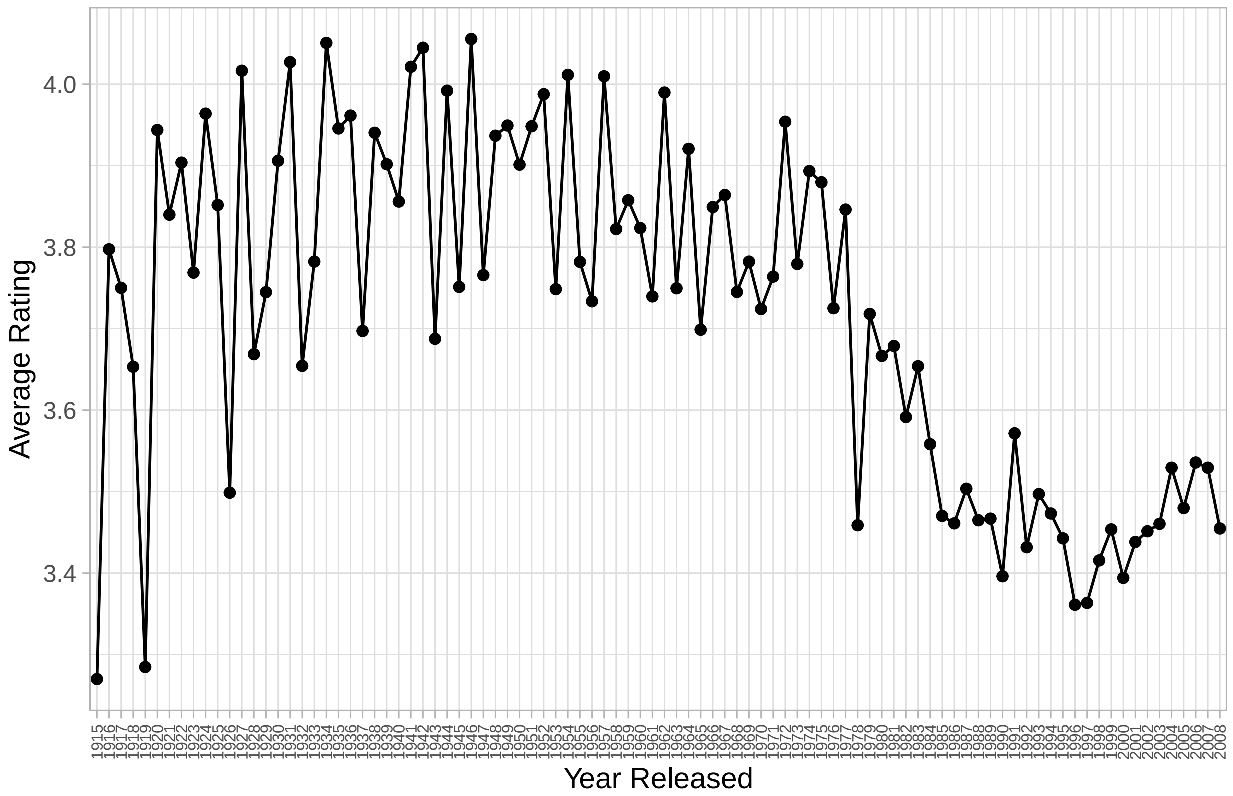


Figure 6. Evolution of Average Rating by Year of Movie Release



2.3 Modeling Approach

In this section of the report, we examine the techniques used in modelling. The accuracy of the recommendation system would be evaluated using the Root Mean Squared Error (RMSE). In simple terms, the RMSE is the standard deviation of the prediction errors. In geometric terms, it measures how concentrated or far the error terms are from the line of best fit. There are many advantages associated with using the RMSE in evaluating recommendation system. First, it is easily understood since it is in the same unit as the original dataset. Next, it is a mathematically convenient method in evaluating distance and gradient metrics, making it preferred to alternative evaluation metrics such as the Mean Average Error (MAE). Additionally, the RMSE tend to penalize large errors, thus could easily help us detect the consequences of outlier predictions in our methods. The RMSE is given by the formula below:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}$$

where Y_i is the actual i-th movie rating, $\hat{Y}_i$ is the predicted rating and N is the number of observations. The best model should have the lowest RMSE among models with RMSE < 0.86490. The RMSE metric was represented in coding by the following:

```
RMSE <- function(true_ratings, predicted_ratings){
  sqrt(mean((true_ratings - predicted_ratings)^2))
}
```

2.3.1 Method 1: The Overall Mean

The first method to be examined is the overall average rating of the movie by users irrespective of any other characteristics. In this model, we assume that ratings of all users tend to a particular average and that differences in individual ratings represent random variations. Consequently, the model assumes that the expected value of ratings equals the overall mean rating. The overall mean model can be represented in the formula below:

$$Y_{u,i} = \mu + \varepsilon_{u,i}$$

where $Y_{u,i}$ is the i-th movie rating of user u, μ is the overall mean and $\varepsilon_{u,i}$ is the error term.

From the estimation, the overall mean rating of the train set is 3.5124556 and the corresponding RMSE on the test data is 1.060054.

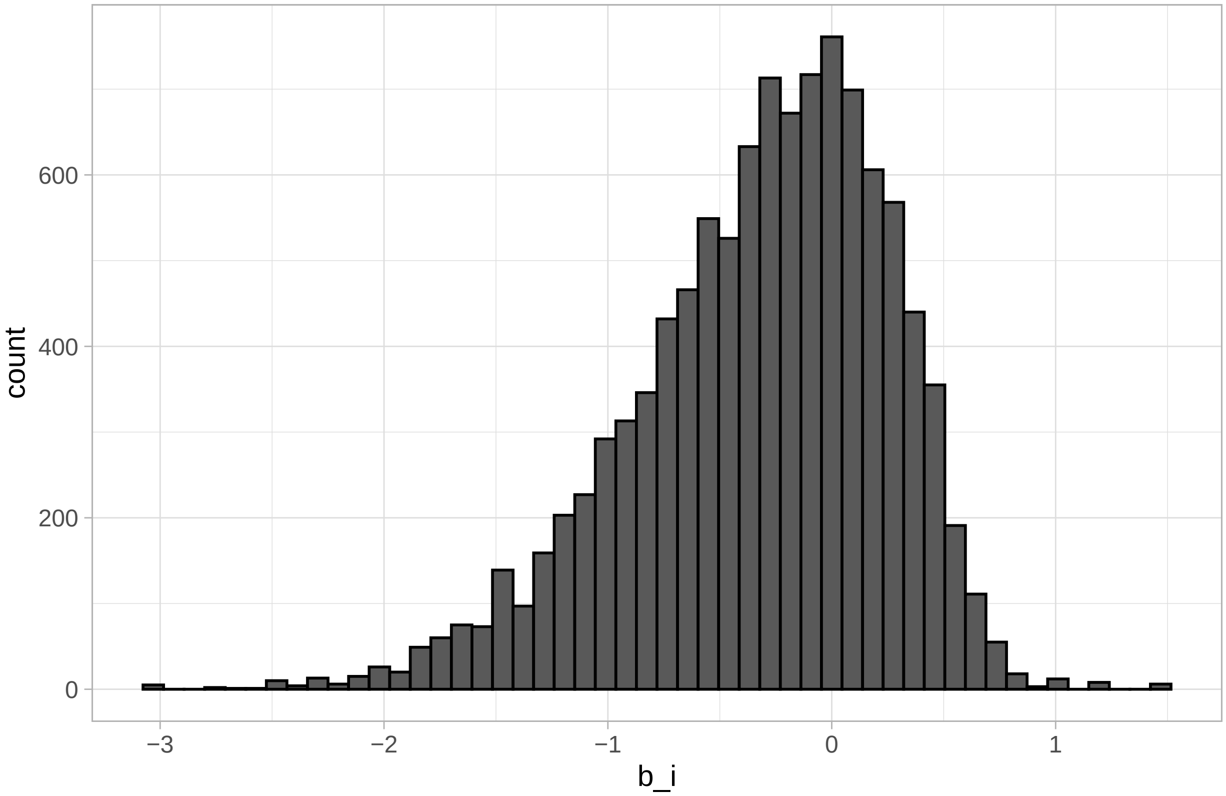
2.3.2 Method 2: Incorporating Movie Effects

The perceived quality of movies and their associated ratings differ from movie to movie. Therefore, users would rate movies differently; an indication that there are movie effects in rating. Hence, the prior method that did not consider movie effects is likely biased. Incorporating the movie effects can be represented in the model below:

$$Y_{u,i} = \mu + b_i + \varepsilon_{u,i}$$

where $Y_{u,i}$ is the i-th movie rating of user u, μ is the overall mean, $\varepsilon_{u,i}$ is the error term and b_i is the movie effect.

Figure 7. Movie Effect



The approximated estimated residual for this model is given as

$$\varepsilon_{i,u} = y_{u,i} - \hat{\mu}$$

The RMSE of the model with movie effects incorporated is 0.942961.

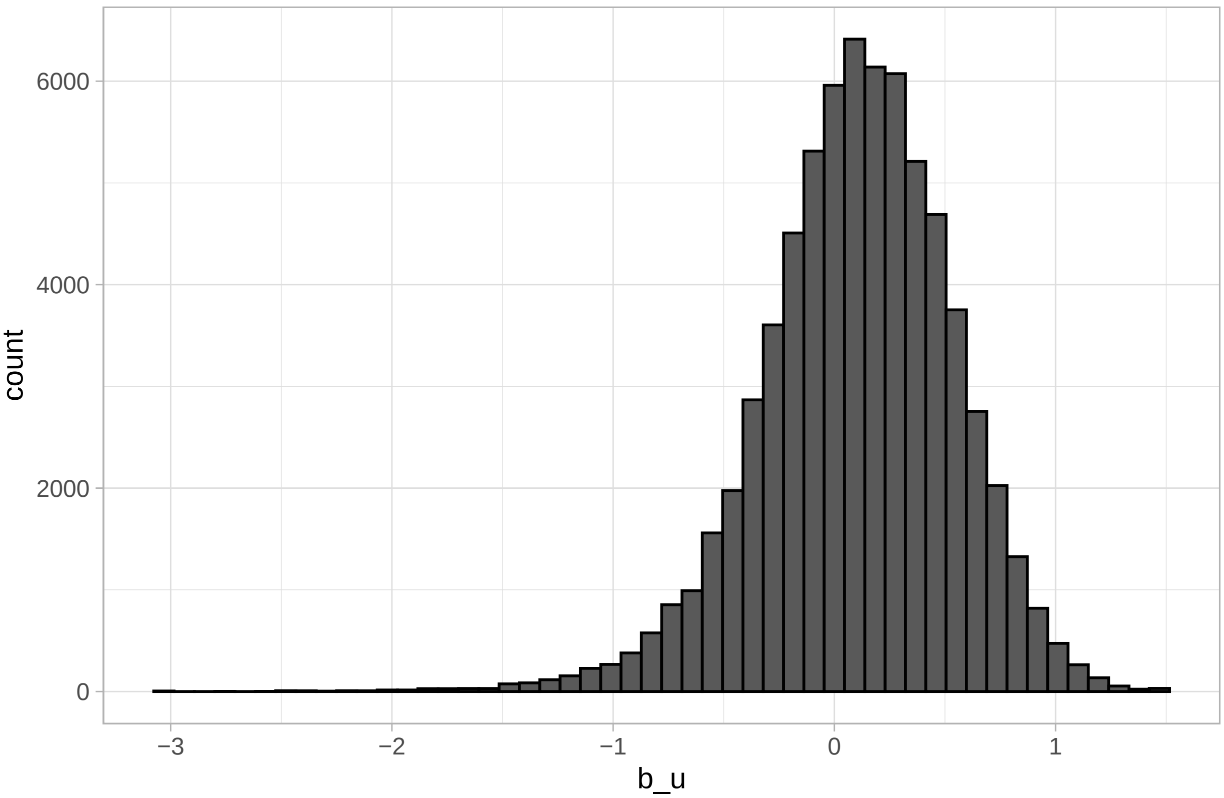
2.3.3 Method 3: Incorporating User Effects

Not surprisingly, the ratings of users tend to differ from user to user. In other words, there are differences in user ratings owing to several inherent characteristics. As regard to any form of utility, individual motivations for ratings are varied and based on multiple factors. For instance, some users are very critical whiles others are liberal. Also, some users are technical in their approach to movie ratings whiles others rate movies from non-technical lenses. Irrespective of reason for specific ratings, users tend to be unique in their ratings. The uniqueness of user ratings is incorporated in the previous model and represented below:

$$Y_{u,i} = \mu + b_i + b_u + \varepsilon_{u,i}$$

where $Y_{u,i}$ is the i-th movie rating of user u, μ is the overall mean, b_i is the movie effect, b_u is the user effect and $\varepsilon_{u,i}$ is the error term.

Figure 8. Individual User Effects



The approximated estimated residual for this model is given as

$$\varepsilon_{u,i} = y_{u,i} - \hat{\mu} - \hat{b}_i$$

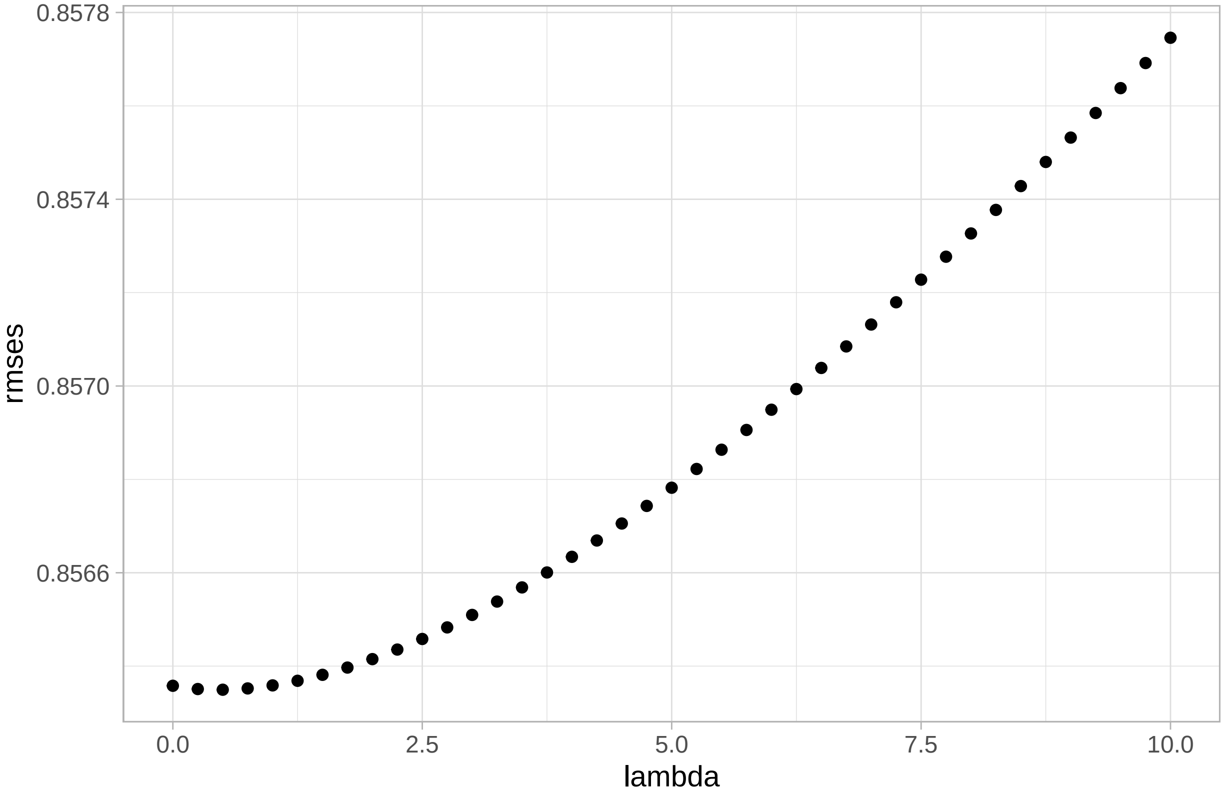
The RMSE of the model with movie and user effects incorporated is 0.864684.

2.3.4 Method 4: Regularization

Thus far, we have noted that the number of user ratings per movie is not same, some movies are rated more than others. Therefore, there is the tendency for rare movies with few good or bad reviews to have very high or bad average reviews respectively, especially when rated by critical or liberal reviewers. Additionally, popular movies with many reviews would generally tend to have mid-level ratings because there is the high likelihood of such movies receiving ratings from both non-technical and technical audiences. The method of regularization incorporates the number of reviewers per movie into the movie rating predicting, penalizing the ratings of movies with few ratings. A basic assumption herein is that large errors do increase RMSE and we would accept to be conservative when unsure about true ratings. One important component of the model with regularization is to choose an appropriate penalty term (λ) through minimizing the below:

$$\frac{1}{N}\sum_{u,i}(y_{u,i}-\mu-b_i-b_u)^2+\lambda(\sum_i b_i^2+\sum_u b_u^2)$$

Figure 9. Regularization Method: Plot of RMSES and Lambda



Using cross-validation, the plot above shows the various combinations of lambdas and their corresponding RMSEs. The optimal penalty term is 0.5.

The RMSE on movie regularization and user effects result in an RMSE of 0.864552 on the test set, not much improvement from the previous method. The result show that the few ratings for some movies did not impact on the predicted ratings that much in this data. Therefore, the latter method would exclude regularization to reduce further complexity.

2.3.5 Method 5: Modelling Evidence of Time Effects

At this stage, the movie year time effects were incorporated into the model to understand how the year of movie release and year of movie rating affected actual rating. Prior, the plot of movie release and movie rating had shown variation across time periods. The contributions of year of movie release and year of movie ratings (without regularization) changes the model to the following:

$$Y_{u,i} = \mu + b_i + b_u + b_{yreleased} + b_{yrated} + \varepsilon_{u,i}$$

where $Y_{u,i}$ is the i-th movie rating of user u, μ is the overall mean, b_i is the movie effect, b_u is the user effect, $b_{yreleased}$ is year of movie release, b_{yrated} is year of movie rated and $\varepsilon_{u,i}$ is the error term.

Figure 10. Incorporating Year Movie was Released

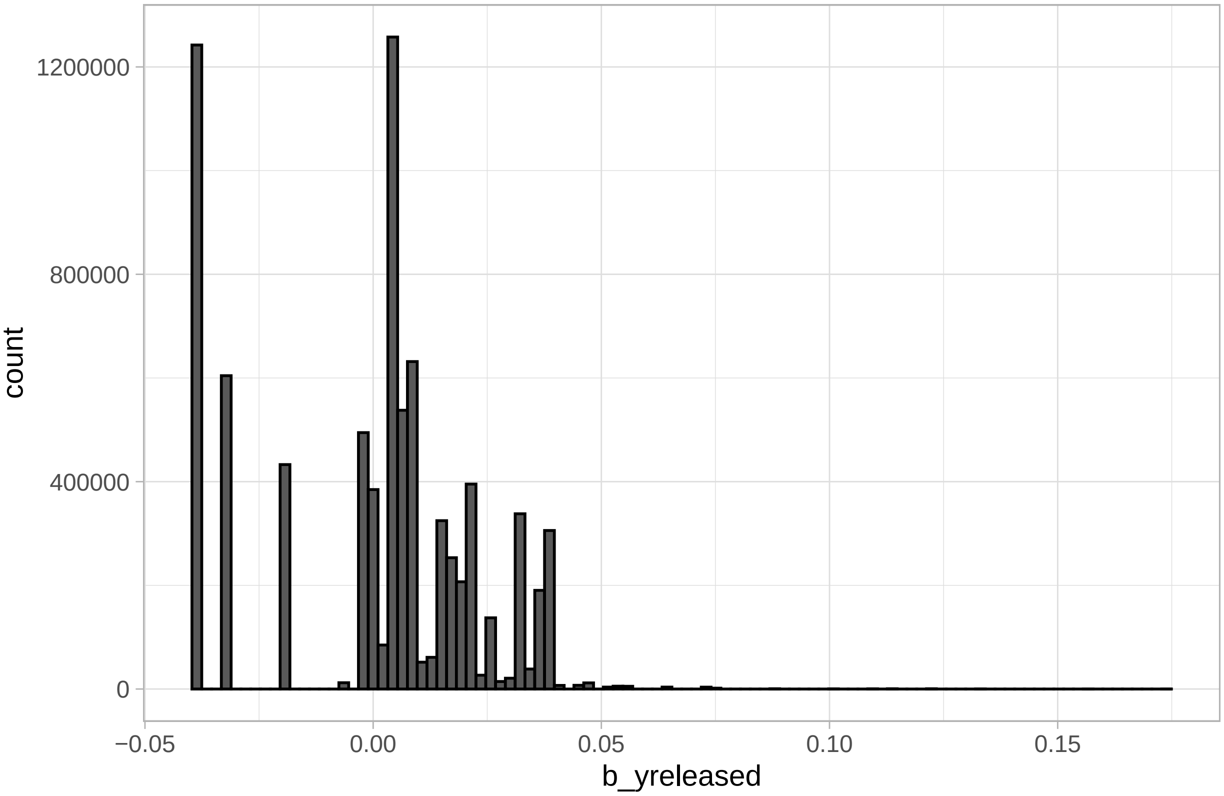
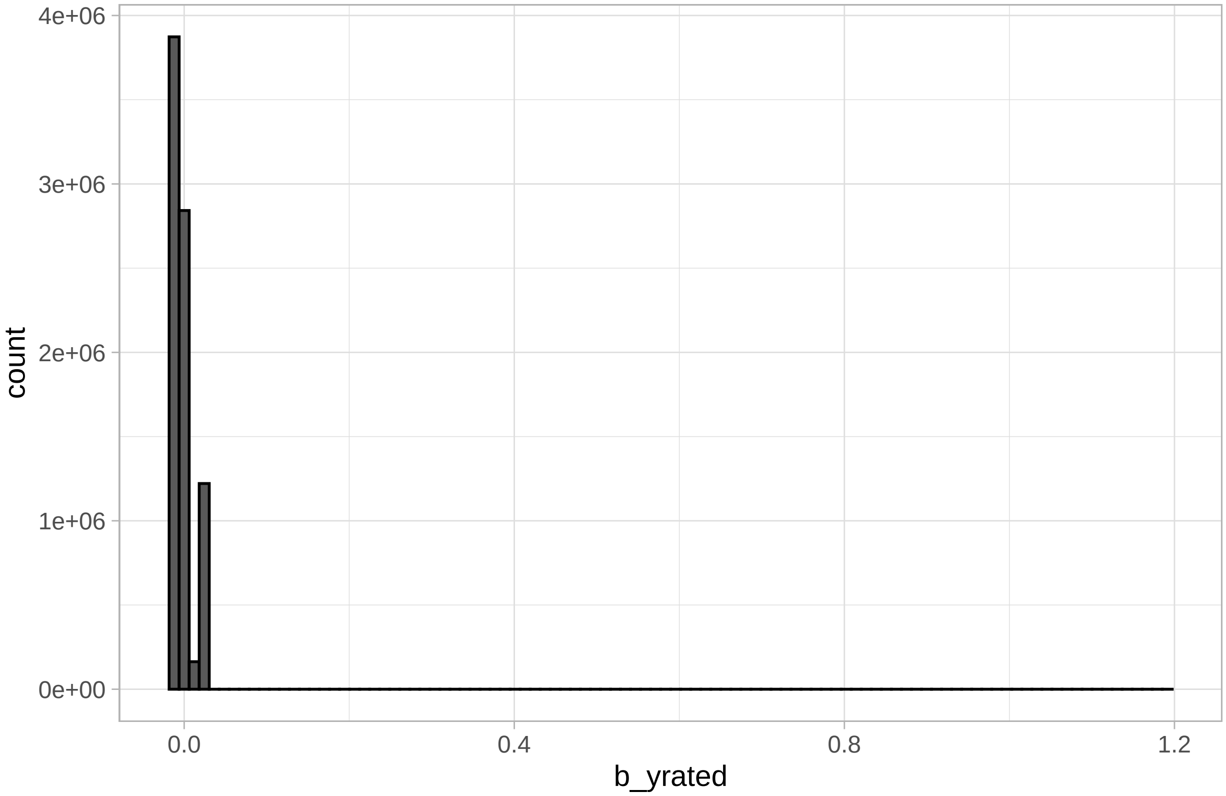


Figure 10 and Figure 11 shows the distribution of subsequently incorporating the year movie was released and the year movie was rated in the model. Both plots show that most observations are centered around zero (0), though there exist skewness. Figure 11 exhibits much skewness to the positive, an indication that there is tendency for ratings when accounted for other properties such as movie and user effects to be at least at the mid-scale. This evidence of the data throw more light on Figure 1, which shown that individuals are likely to rate movies favorably.

Figure 11. Incorporating Year Movie was Rated



The RMSE of the resulting model at this stage on the test set is 0.864273.

2.3.6 Method 6: Time Interaction Effect

In here, we examine the improvement of the model by incorporating an interaction term effect between the year of movie released and the year of movie rated. Specifically, the model proposes that movie rating can differ based on the combination of year release and year of rating. The contribution of the interaction effect changes the model to the following:

$$Y_{u,i} = \mu + b_i + b_u + b_{yreleased} + b_{yrated} + b_{yreleased}b_{yrated} + \varepsilon_{u,i}$$

where $Y_{u,i}$ is the i-th movie rating of user u, μ is the overall mean, b_i is the movie effect, b_u is the user effect, $b_{yreleased}$ is year of movie release, b_{yrated} is year of movie rated, $b_{yreleased}b_{yrated}$ and $\varepsilon_{u,i}$ is the error term.

The tables below show the average rating of the top 10 frequent and 10 non-frequent interaction term.

Year Released Year Rated	Frequency	Average Rating
1995_1996	267221	3.5031
1994_1996	248682	3.4982
1993_1996	139213	3.5640
1996_1997	131775	3.4233
1999_2000	93256	3.4859
1998_2000	70644	3.4587
1997_2000	67857	3.4118
1996_1996	66630	3.5382

Year Released Year Rated	Frequency	Average Rating
1995_1997	66268	3.5325
2000_2001	64761	3.4588

Year Released Year Rated	Frequency	Average Rating
1926_2009	1	1.0
1923_2009	1	3.0
1925_2009	1	3.0
1918_2005	1	3.5
2002_2001	1	4.0
1920_2000	1	4.0
1918_2006	1	4.5
1995_1995	1	5.0
2006_2005	2	2.5
1920_2002	2	3.0

The top 10 most frequent combination of year released/year rated movies have ratings between 3.41 and 3.57. However, for the non-frequent interaction, the average rating ranges from 1.0 to 5.0, though the modal rating is 3.0. Due to the inclusion of the interaction term, there could be the possibility of predictions that could result in missing values or outside the actual rating bound of 0.5 and 5.0. Specifically, only one prediction came out as missing value and was imputed by the average value of predicted ratings which has been shown to be similar with both the top 10 frequent and 10 non-frequent interaction terms shown in the tables above. Also, predicted ratings above the rating ceiling of 5.0 were set to the ceiling whiles the predicated ratings below the rating floor of 0.5 were set to the rating floor.

The RMSE of the resulting model at this stage on the test set is 0.863374.

3. Results

In the results section, we re-examine the modeling results and discuss the model performance. The final model resulted in the lowest RMSE on the test and would be evaluated on the validation set.

The results of all the methods are presented below.

Method	RMSE
Just the Average	1.060054
Movie Effect Model on Test Set	0.942961
Movie and User Effects Model on Test Set	0.864684
Reg. Movie and User Effects Model on Test Set	0.864552
Movie, User, Year Released and Year Rated Effects Model on Test Set	0.864273
Movie, User, Year Released, Year Rated and Year Interaction Effects Model on Test Set	0.863374
Movie, User, Year Released, Year Rated and Year Interaction Effects Model on Validation Set	0.864528

As noted earlier, preferred models should have an RMSE below 0.86490 on the validation set. From the table above, we note that the lowest RMSE on the test set was 0.863374. Thus, the chosen model is the final model, the model that involved the inclusion of the interaction term. Moreover, the chosen model performed very well on the validation set, producing an RMSE of **0.864528**, similar to the RMSE on the test data and less than the threshold of 0.86490. Therefore, the result is robust to either over-fitting or under-fitting. It is expected that the model will still perform well on another set of new data.

4. Conclusion

In this report, we examined model techniques in predicting movie ratings using RMSE as the metric of evaluation. The data used to train the model comprised 69878 unique users rating 10677 unique movies, an indication that number of movie ratings were unequal among users. However, regularization technique shown that accounting for the unequal numbers did not account for significant bias in training. Overall, the chosen model involved the inclusion of interaction time effects of year of movie release and year of movie rated, after accounting for user, movie, year of movie release and year of movie rated. The model produced an RMSE of 0.864528 on the validation set, lower than the threshold of 0.86490.

That notwithstanding, we acknowledge some limitations. First, more advanced algorithms such as matrix factorization could produce superior RMSE. However, current matrix factorization could be computationally expensive and time consuming. Also, movie ratings can be thought of as a form of utility maximization, where ratings are used to express satisfaction. In this, utility studies suggest that taste changes, making it possible for the same users to rate the same movie differently under different circumstances. In effect, the data used in the estimation was obtained at a point in time in the past which may not be consistent over time. Again, the data does not tell us how the ratings were obtained. This is important because it is possible for non-humans (computers) to provide ratings, borne out of marketing techniques. Such artificial ratings if prevalent could undermine the modelling of human user ratings.

Recommendation systems tend to be important component of building successful businesses nowadays. Future work could expand on computationally inexpensive but advanced algorithms and economic motivations of users in their rating decisions. Additionally, future research could also incorporate indications of general locality to determine whether there are geographical effects. In all analysis, it is recommended that obvious individual identifier information are limited to the barest minimum.

Reference Irizarry, R.A. (2022).Introduction to Data Science: Data Analysis and Prediction Algorithms with R. *The coding in this report was built on primary code provided by the course as contained in Irizarry(2022).*