Movie Recommendation System

Hugh Morris

27/11/2020

EXECUTIVE SUMMARY

A dataset consisting of approximately 10 million movie ratings for 10,677 movies, from 69,878 users of the movie recommender service MovieLens, was used to construct a movie recommendation system. 90% of the dataset was used for training and tuning model parameters (edx dataset), while the remaining 10% was used to test the accuracy of the model (validation dataset). The goal was to produce a model that has a root mean square error (RMSE) of less than 0.86490

To build the movie recommendation system, information on the average effects of each movie, user , genre, as well as date were used. The date effects captures information on when a movie premiered, the date (year, month, week, day) it was rated, in addition to the gap between when a movie was rated and when it premiered in years. The edx dataset was further subdivided into edx_train, which was used to build the model and edx_test, which was used to tune model parameters (penalty terms), and the one that produced the lowest RMSE chosen.

Exploratory data analysis was undertaken to aid in model development. Six predictors were included in the model, movieId; userId, genres; year premiered; gap between when a movie was rated and when it is premiered (gap effect); and year rated. Given the small sample sizes for some of the predictors, a regularization approach was used in model development.

The regularized model optimizes the effects (movies, users, etc) at each stage of the model construction (that is the penalty term is unique for each variable), which was then used in all subsequent stages. The model was then applied to the validation dataset to predict ratings a user will give a movie and produced a RMSE of 0.86433, which was below the target of 0.86490. It should be noted that if the date effects were not included in the model the RMSE for the validation dataset would have been 0.86491, which is above the desired level.

Key Words: Movie Recommendation System, Regularization, Feature Engineering

SECTION I: INTRODUCTION

The increase in options available to consumers have made recommendation systems a very useful tool in aiding consumers when making decisions. This tool is especially useful in making movie recommendations given the large number of movies available.

The objective of this paper is to develop an algorithm that predicts the rating a user will give to a movie, which will then be used to make recommendations to the user. The data set used in the paper contains, 10,000,054 movie ratings for 10,677 movies, from 69,878 users selected at random with at least 20 movie ratings, from the movie recommender service MovieLens 10M Dataset(Harper and Konstan 2016).

In order to assess the accuracy of the movie recommendation model, the movieLens dataset was subdivided into edx dataset, for training and tuning of model parameters, and a validation dataset to test the accuracy of the model. The goal is to construct a model using the edx data set, which will be used to predict users' rating of movies in the validation dataset, with a root mean square error (RMSE) of less than 0.86490.

To construct the recommendation model, the paper will utilize information on movie effects, user effects, genre effects, as well as, date effects. The date effects seek to capture the effects of when the movie premiered, the date when the movie was rated (year, month and week), in addition to the gap in years between when a movie premiered and when it was rated. The model will be constructed by estimating the average effect of each variable when controlling for the other variables.

The rest of the paper will be as follows:

- Section II will give a brief description of the dataset, feature engineering, data cleaning and exploratory data analysis
- Section III will present the methodology used in constructing the model
- Section IV will present the modeling results and performance
- Section VI will conclude by summarizing key findings, speak to limitations and recommendation for future work.

SECTION II: DATA DESCRIPTION, WRANGLING AND EXPLORATORY DATA ANALYSIS (EDA)

A key step in any data analysis is to first describe the data used in the analysis, undertake data wrangling to create features (predictors) and clean the dataset, as well as undertake exploratory data analysis to inform potential variable selections that should be included in the analysis.

Data Description

The movielens dataset consist of 6 columns with 10,000,054 rows. The columns are as follows:

- userId - represents anonymized user identification. There are 69,878 unique users
- movieId - represents the identification given to each movie by MovieLens. There are 10,677 unique movies
- rating - represents the rating given to a movie by a user. Ratings are made on a 5 point scale with 0.5 points increment starting from 0.5. There are 10 unique ratings
- timestamp - represents the date and time the movie was rated and it is measured in seconds since midnight Coordinated Universal Time (UTC) of Jauary 1, 1970. There are 7,096,905 unique timestamps
- title - represents the movie title and include the year the movie premiered at the end in bracket, there are 10,676 unique titles
- genres - represents a pipe-separated list of different genres. They are 797 unique genre groups

Table 1 below gives the first 5 observations within the movielens dataset. The first thing to note is that the first 5 rows only have information on one user, whose anonymized userId is 1. This user has given the first 5 movies (based on unique movieId and title) a rating of 5. Notice that the movie title has the year the movie premiered in bracket at the end. The genres variable includes all the genres the movie is classified under.

TABLE 1
First 5 rows of MovieLens 10M 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	231	5	838983392	Dumb & Dumber (1994)	Comedy
1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi

The objective is to develop a model to predict the rating a user will give to a movie. Therefore, rating is our dependent variable, and the other variables are potential independent variables that can be used to aid in predicting the rating each user will gives to a movie.

To reduce the likelihood of overfitting, the movielens dataset is randomly subdivided into edx dataset (90% of the movielens dataset) for model training and parameter tuning; and validation dataset (10% of the movielens dataset) which will represent the final holdout test set for which the model will be evaluated. Since the edx dataset will be used to train the model, it was ensured that all the users and movie identifications in the movielens dataset were also in the edx dataset.

Data Wrangling

For the remainder of this section the focus will be on the edx dataset, the validation dataset will be treated as if it is unknown, until it is time to assess the accuracy of our model using the RMSE. The goal is to identify the predictors/independent variables from the remaining 5 variables to build a model. While some of the potential indicators are clear, for example, using the variables movieId, userId and genres, the other potential predictors require some data wrangling to be of use.

The first problem that we identified was that the title column is not tidy, it has the year in which the movie premiered in its title (see Table 1). Secondly, the timestamp can be adjusted to represent date and time, which may be used to generate other predictors, a technique commonly called feature engineering (Kuhn and Johnson 2019).

The following adjustments were made to the edx dataset:

- the year the movie premiered was extracted from the title column and called year_movie

TABLE 2a
First 5 rows of title and year_movie from edx dataset

title	year_movie
Boomerang (1992)	1992
Net, The (1995)	1995
Outbreak (1995)	1995
Stargate (1994)	1994
Star Trek: Generations (1994)	1994

- the timestamp column was converted into a date and time format, thereafter, the year, month , week and weekday in which the movie was rated was extracted as potential predictors. It should be noted that the hour, minutes and second could have been extracted as potential predictors.
  - year_rated, which ranges from 1915 to 2008
  - month_rated, which ranges from 1 (January) to 12 (December)
  - week_rated, which ranges from 1 (first week) to 53 (last week in a leap year)
  - day_rated, which ranges from 1 (Sunday) to 7 (Saturday). For example, the 2 of August in 1996, is a Friday, which is the sixth day of the week (Table 2b).

TABLE 2b
First 5 rows of timestamp and its features from edx dataset

timestamp	date	year_rated	month_rated	week_rated	day_rated
838985046	1996-08-02 11:24:06	1996	8	31	6
838983525	1996-08-02 10:58:45	1996	8	31	6
838983421	1996-08-02 10:57:01	1996	8	31	6
838983392	1996-08-02 10:56:32	1996	8	31	6
838983392	1996-08-02 10:56:32	1996	8	31	6

- The gap between when a movie premiered and when it was rated in years is also included as an additional predictor. That is, year_movie - year_rated = year_gap. The assumption is that the gap between when a movie premiered and when it is rated would influence the rating it receives. This relationship may be negative, as the gap between when a movie premiered and when it is rated increases, it is hypothesized that the technology used to produce the movie in the past would be below the standard of the time the movie was rated the larger the gap year, so users are more likely to rate movies that has a large gap year less favorable (*technology effect*).
- The relationship between gap year and average rating may also be positive, as some users are more likely to rate a movie that has a large gap year higher because of the *nostalgia effects*. The view that things made in the past are better.

TABLE 2c

First 5 rows of title, year_movie, year_rated and year_gap from edx dataset

title	year_movie	year_rated	year_gap
Boomerang (1992)	1992	1996	4
Net, The (1995)	1995	1996	1
Outbreak (1995)	1995	1996	1
Stargate (1994)	1994	1996	2
Star Trek: Generations (1994)	1994	1996	2

Now that we have created some features, the edx dataset is randomly split into edx_train, which accounts for 90% of the edx dataset and edx_test which account for the remaining 10% of the edx dataset. The model will be built on edx_train with its parameters calibrated/tuned using the edx_test dataset. The datasets were adjusted to ensure that edx_train has all the movies and user ids in the edx dataset.

The next step in the process is exploratory data analysis, which will inform data cleaning and variables to include in model (Figure 1). Exploratory data analysis and data cleaning goes hand in hand and will impact how features are generated.

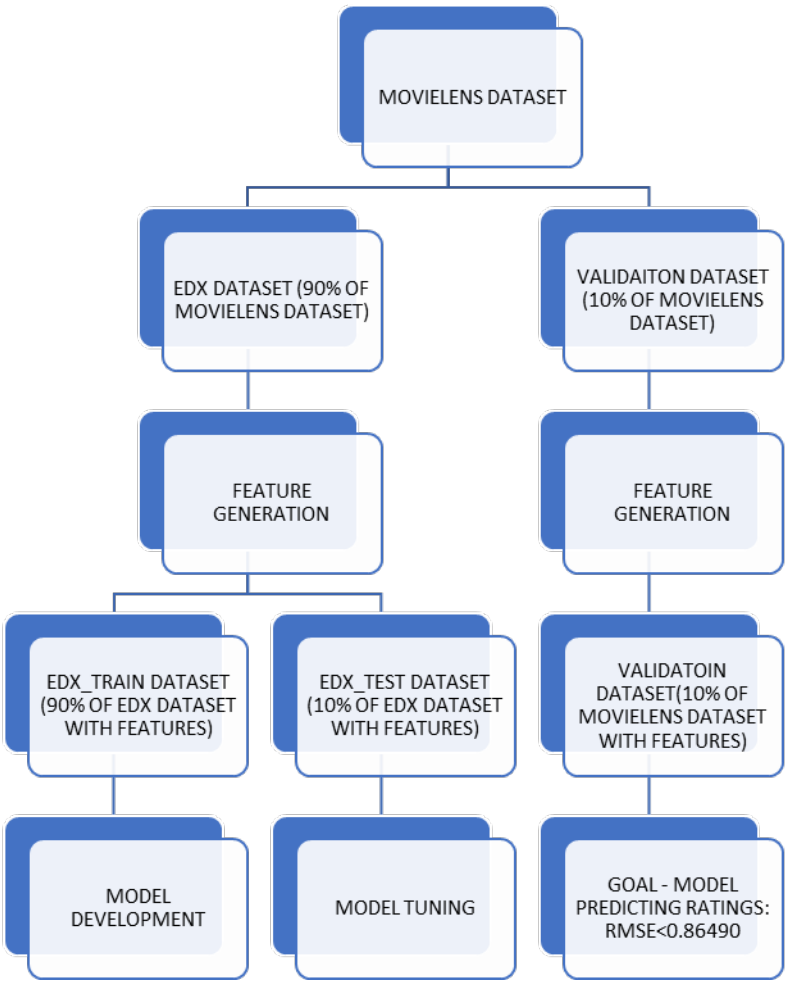


FIGURE 1: WORK FLOW PROCESS

Exploratory Data Analysis

One of the tools used to identify variables to include and the modeling approach to use is exploratory data analysis, where graphical analysis is used to understand the underlying data structure. Each variable will be analyzed by itself and with respect to the dependent variable. Exploratory data analysis will be undertaken on the edx_train dataset.

Rating

The rating variable represents the rating of the 'm' movie, by the 'u' user, and ranges from 0.5 to 5.0, with a mean of 3.51 and median of 4.0, which means that the variable is negatively skewed. The most frequent number (mode) is 4.0. Additionally, a user is more likely to rate a movie as a whole number, than a rating that ends with 0.5 (Table 3 and Figure 2).

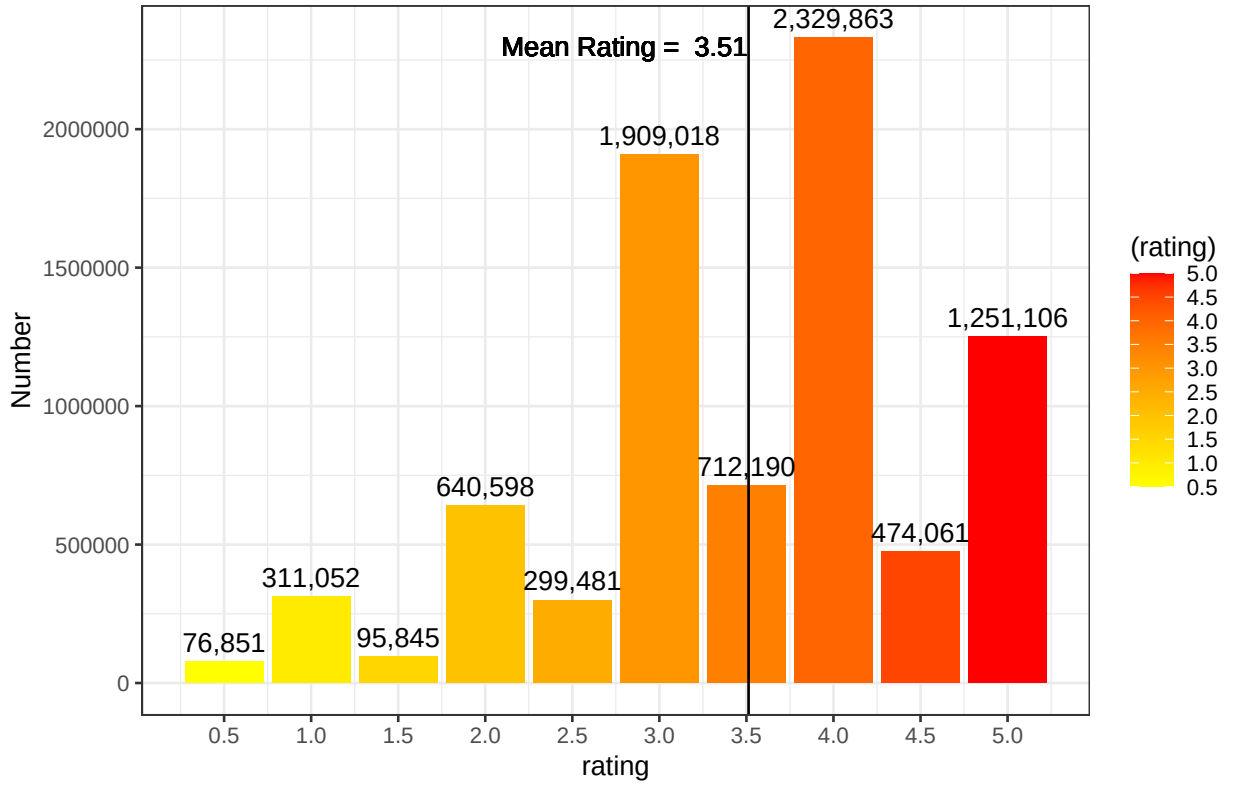


FIGURE 2: NUMBER OF RATINGS FOR EACH RATING CATEGORY

TABLE 3
DISTRIBUTION OF RATING VARIABLE

INDICATORS	ESTIMATES
minimum (min)	0.500
first_quartile (Q1)	3.000
median	4.000
mean	3.512
third_quartile (Q3)	4.000
maximum (max)	5.000
standard_deviation	1.060

MovieId

The movieId variable is the unique id used to identify movies. There are 10,677 unique movie id in the edx_train dataset. It is a potentially useful indicator in constructing a model to predict the ratings a user will give a movie. Some movies are of a high quality and therefore are expected to be rated highly by most users on average, while some movies are of low quality and thus expected to be rated low by users on average. This effect we will call the movie effect. The reliability of this effect depends on the number of users that rated the movie, for example, if only one person rated the movie a 5 star, then we would be less confident to predict a 5 star for that movie in general, compared with if 1,000 persons on average rated the movie 5 star. Therefore it is not only important to know how high the movie was rated on average, but also how many users contributed to this average rating.

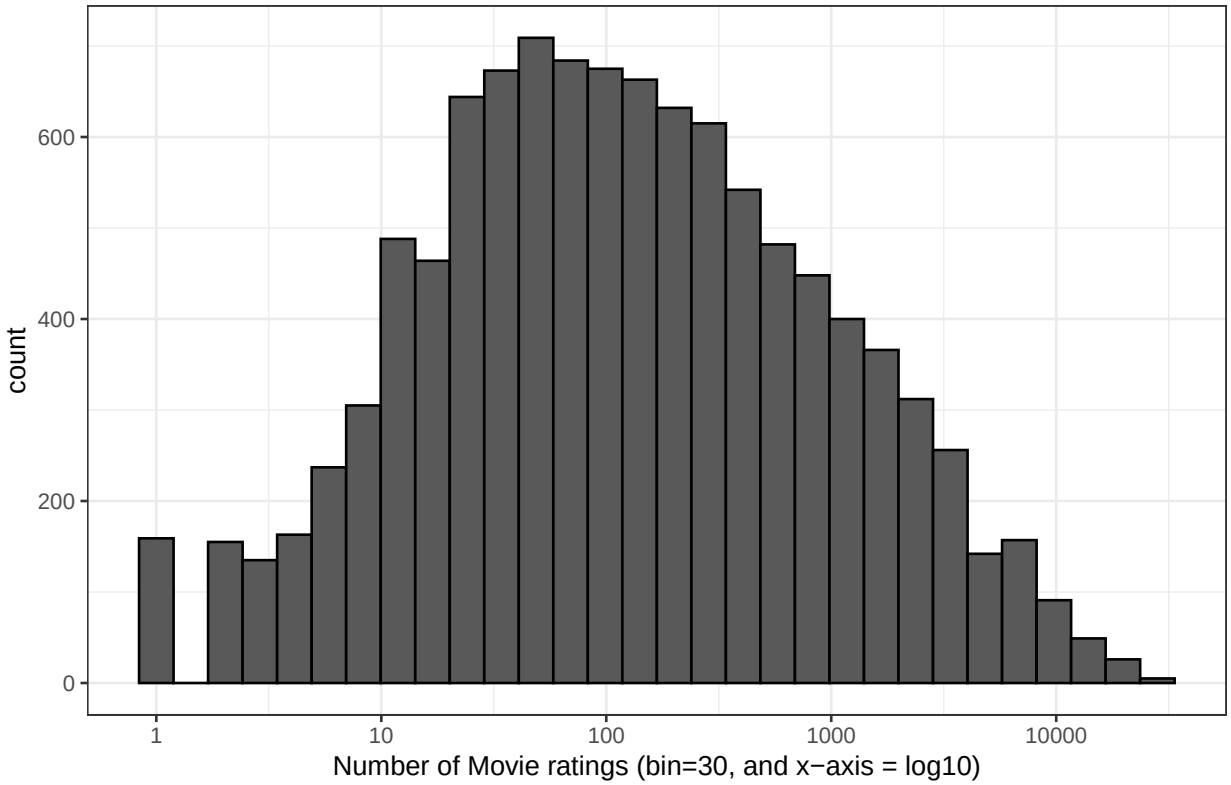


FIGURE 3: NUMBER OF MOVE RATINGS BY MOVIE ID

A graphical display of the number of times a movie was rated revealed that some movies were rated frequently, while others were rated infrequently (Figure 3). This was because some movies are blockbusters, for example, movieId 296, is for the movie entitled ***Pulp Fiction*** which is rated the most (28,168) in the dataset with an average rating of 4.154 out of 5, which suggests that ***Pulp Fiction*** is a high quality movie (Table 4a).

TABLE 4a
TOP 10 MOST RATE MOVIES

movieId	title	count	avg_rating
296	Pulp Fiction (1994)	28,168	4.154
356	Forrest Gump (1994)	27,987	4.013
593	Silence of the Lambs, The (1991)	27,327	4.203
480	Jurassic Park (1993)	26,381	3.666
318	Shawshank Redemption, The (1994)	25,188	4.457
110	Braveheart (1995)	23,545	4.082
457	Fugitive, The (1993)	23,397	4.011
589	Terminator 2: Judgment Day (1991)	23,332	3.926
260	Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	23,065	4.221
592	Batman (1989)	21,857	3.383

Similarly, less known movies were rated infrequently, which means we have less confidence in using the average rating of the movie as an indication of how other users will rate these movies (Table 4b). There are 159 movies with only one rating and 1,154 movies with less than 10 ratings. If we include movieId as a predictor when predicting ratings, the training model will overfit on movie Ids that have a small amount of ratings in the training dataset, and thus give a poor prediction in a dataset not used in the training (validation dataset). For example, movieId 3226, entitled ***Hellhounds on My Trail*** *was rated by one user, who gave it a rating of 5. If a linear model was used to predict the rating of movieId 3226, it will most likely*

overestimate how other users will rate **Hellbounds on My Trail**.

TABLE 4b
TOP 10 LEAST RATED MOVIES)

movieId	title	count	avg_rating
604	Criminals (1996)	1	1.000
3191	Quarry, The (1998)	1	3.500
3226	Hellhounds on My Trail (1999)	1	5.000
3234	Train Ride to Hollywood (1978)	1	3.000
3277	Beloved/Friend (a.k.a. Amigo/Amado) (Amic/Amat) (1999)	1	3.000
3312	McCullochs, The (1975)	1	2.000
3356	Condo Painting (2000)	1	3.000
3376	Fantastic Night, The (La Nuit Fantastique) (1942)	1	3.000
3383	Big Fella (1937)	1	3.000
3561	Stacy's Knights (1982)	1	1.000

To control for overfitting because of small sample size, one method often used in the literature is regularization, which seeks to shrink or regularize the coefficients, which reduces the variance, with a negligible increase in bias (James et al. 2013).

Another important consideration is the variability of the indicator, that is, can we find significant differences in the ratings of movies, or, are most movies rated the same in general. The average movie ratings for each movie showed that some movies are rated low (0.5) on average, while others are rated high(5.0) on average (Table 4c). These average ratings however have to be viewed with caution given the number of times that the movie has been rated ranges from 1 to 28,168 (see Table 4b and Table 4c) .

TABLE 4c
DISTRIBUTION OF AVERAGE MOVIE RATINGS BY MOVIEID

Indicator	movieId_avg_rating
min	0.500
Q1	2.842
median	3.269
Q3	3.613
max	5.000
Inter-Quartile Range	0.771
range	4.500

UserId

The userId variable, which represents the unique id for each user, is also a potential useful predictor when predicting ratings. There are 69,878 users in the edx_train dataset. It is theorized that certain users are more generous in their ratings (i.e. more likely to rate a movie above its average rating), while others are more frugal (i.e. more likely to rate a movie below its average rating), this we will call the user effect. The confidence in which we can say a user is generous or frugal will be dependent on the extent to which they have rated a sufficiently large number of movies.

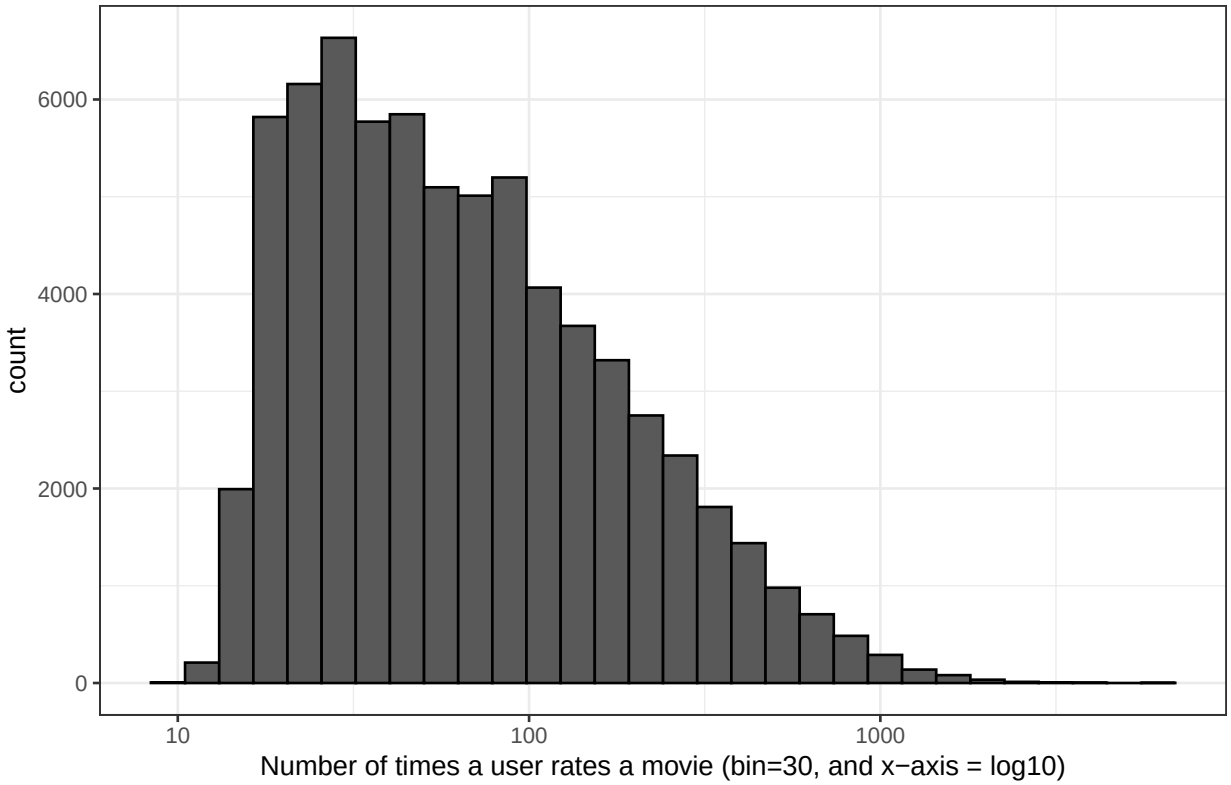


FIGURE 4: NUMBER OF MOVE RATINGS BY USER ID

Figure 4 displays the distribution for the number of times a user rates a movie, similar to the movie Id variable, we have some users that are very active in rating movies, while other users have only rated a few movies.The minimum number of ratings undertaken by a user in the edx_train dataset was 9, with the largest 5,931.The average user effect by userId ranges between 0.5 and 5.0, with the interquartile range being 0.549. (Table 5).

TABLE 5
DISTRIBUTION OF AVERAGE MOVIE RATINGS BY USERID

Indicator	UserId_avg_rating
min	0.500
Q1	3.356
median	3.634
Q3	3.905
max	5.000
Inter-Quartile Range	0.549
range	4.500

Genres

The genre variable is used to classify the different style of movies, some movies have several genres, while others fall into one genre. Within the edx_train dataset there are 797 genre groups, which includes individual genres or a combination of genres. Within the genres variable there are 18 genres, IMAX (i.e. high-resolution films) and one movie with no genre listing which was watched 7 times. Majority of the movies in the edx_train dataset had drama, comedy and action in the genres variable, either individually, or as part of a broader genres group (Table 6).

TABLE 6
BASIC STATS ON AVERAGE MOVIE RATINGS AND NUMBER OF RATINGS BY USERID

genres	count
Drama	3,518,986
Comedy	3,187,285
Action	2,304,395
Thriller	2,093,191
Adventure	1,717,992
Romance	1,541,727
Sci-Fi	1,207,309
Crime	1,194,740
Fantasy	833,130
Children	664,146
Horror	622,183
Mystery	511,788
War	460,203
Animation	420,349
Musical	389,569
Western	170,406
Film-Noir	106,704
Documentary	83,664
IMAX	7,358
(no genres listed)	7

The genres group with the highest average rating was Animation|IMAX|Sci-Fi with an average rating of 4.67 (Table 7). However, it should be noted that movies that fall within this genre group was only rated 6 times.

TABLE 7
TOP 10 AVERAGE BY GENRES GROUP

genres	avg_rating	count
Animation|IMAX|Sci-Fi	4.667	6
Drama|Film-Noir|Romance	4.304	2,693
Action|Crime|Drama|IMAX	4.300	2,095
Animation|Children|Comedy|Crime	4.279	6,418
Film-Noir|Mystery	4.239	5,431
Crime|Film-Noir|Mystery	4.224	3,650
Film-Noir|Romance|Thriller	4.217	2,190
Crime|Film-Noir|Thriller	4.211	4,365
Crime|Mystery|Thriller	4.201	24,173
Action|Adventure|Comedy|Fantasy|Romance	4.197	13,329

The genre group with the worst ratings on average was Documentary|Horror, with an average rating of 1.44 (Table 8).

TABLE 8
LOWEST 10 AVERAGE BY GENRES GROUP

genres	avg_rating	count
Documentary|Horror	1.441	547
Action|Horror|Mystery|Thriller	1.614	289
Comedy|Film-Noir|Thriller	1.647	17
Action|Drama|Horror|Sci-Fi	1.750	4
Adventure|Drama|Horror|Sci-Fi|Thriller	1.796	196

Adventure|Animation|Children|Fantasy|Sci-Fi	1.900	627
Action|Adventure|Drama|Fantasy|Sci-Fi	1.902	51
Action|Horror|Mystery|Sci-Fi	1.921	19
Action|Children|Comedy	1.932	460
Action|Adventure|Children	1.934	745

Given the differences in the rating by genres, the genres is also another indicator that can provide insight on how a movie is rated, as a combinations of different genres may have a high rating on average, while other combinations may have relatively low ratings. Table 9 presents the summary statistics of the genre in terms of average ratings. The average ratings by genres ranges from 1.441 to 4.667, with an interquartile range of 0.588 (Table 9).

TABLE 9
DISTRIBUTION OF AVERAGE MOVIE RATINGS BY GENRES

Indicator	genres_avg_rating
min	1.441
Q1	3.112
median	3.439
Q3	3.700
max	4.667
Inter-Quartile Range	0.588
range	3.225

Date Effects

Features were generated using information on the year the movie premiered, and the date the movie was rated (year, month, week and day of the week), as well as the difference between when a movie was rated and when it premiered as potential predictors to be included in the movie recommendations.

Each date feature will be assessed graphically, by looking on the average effect by date period, a 95% error bar that gives us an idea of how confident we should be above the average effect for each date effect, and a local polynomial regression fit (loess) to identify trends. Key takeaways from each date feature within the edx_train dataset are:

Premiere Year

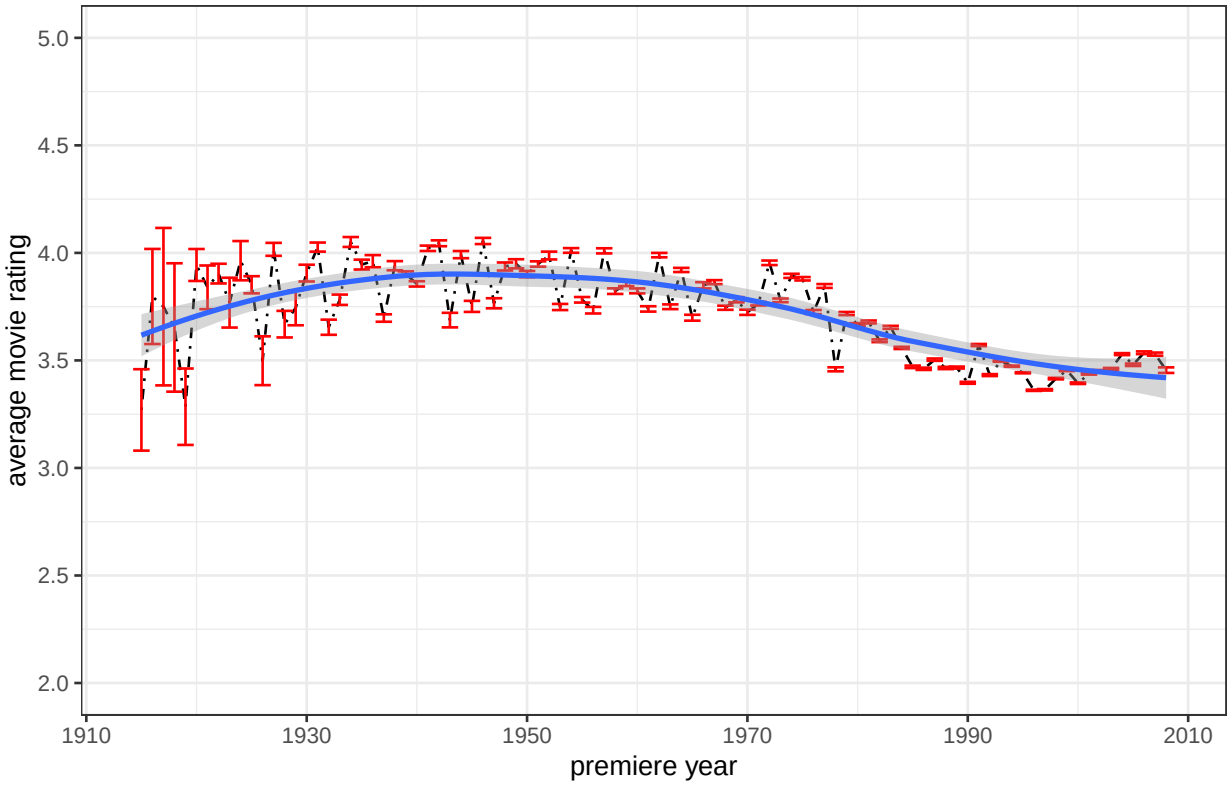


FIGURE 5: AVERAGE EFFECT BY YEAR MOVIE PREMIERED

- The first movie premiered in 1915 and the last movie premiered in 2008.
- There is a upward trend in the average rating over time up to mid 1940s and thereafter the rating based on the year the movie premiered begins to decline.
- Movies that premiered in the earlier periods have wider confidence bands, suggesting higher variability in the ratings in the earlier years, as users are less likely to watch movies in the earliest periods (1915-1930).

year_rated

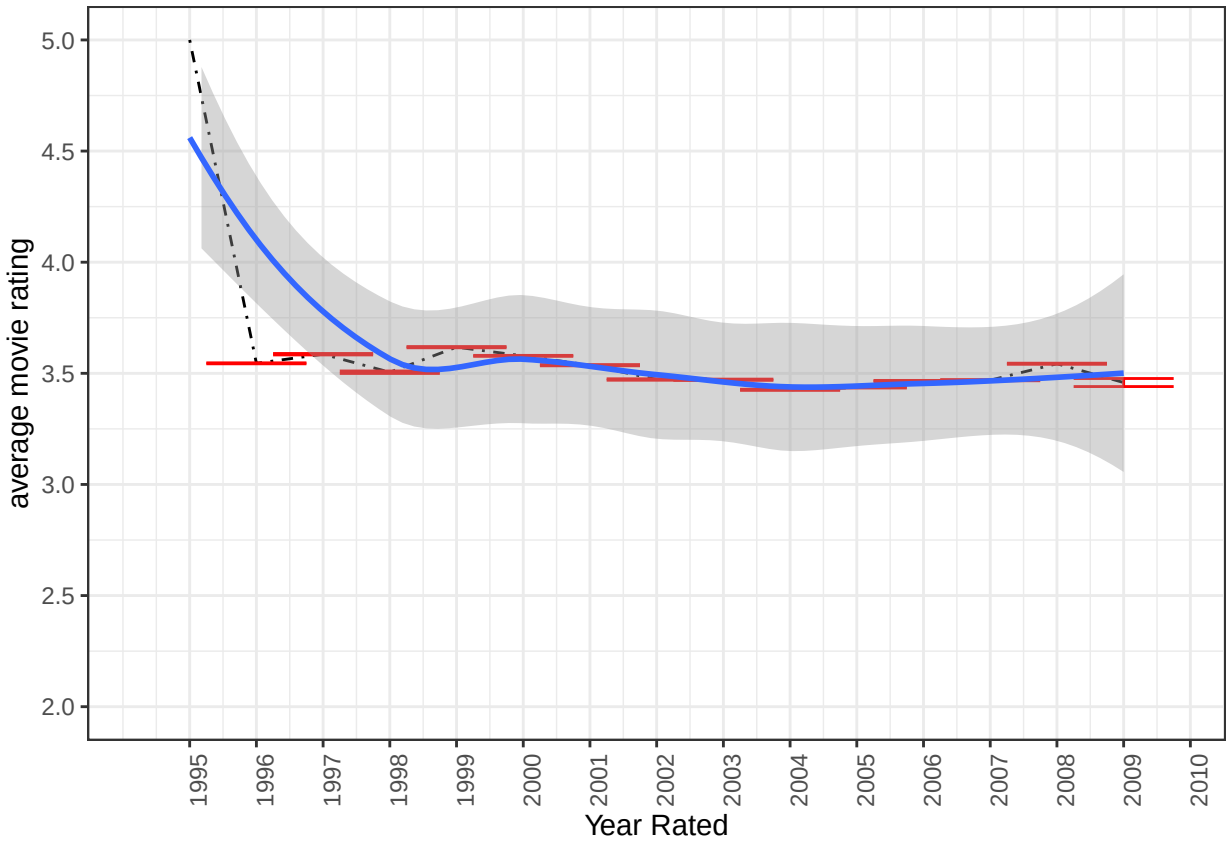


FIGURE 6: AVERAGE EFFECT BY YEAR MOVIE WAS RATED

- The first rating was done in 1995 and the last rating was done in 2009.
- 1995, does not have a 95% confidence band, as only one person rated a movie in 1995. Therefore the rating for 1995 should be interpreted with caution.
- The average rating for movies rated in 2001 and earlier, were higher than movies rated after 2001.

year_gap

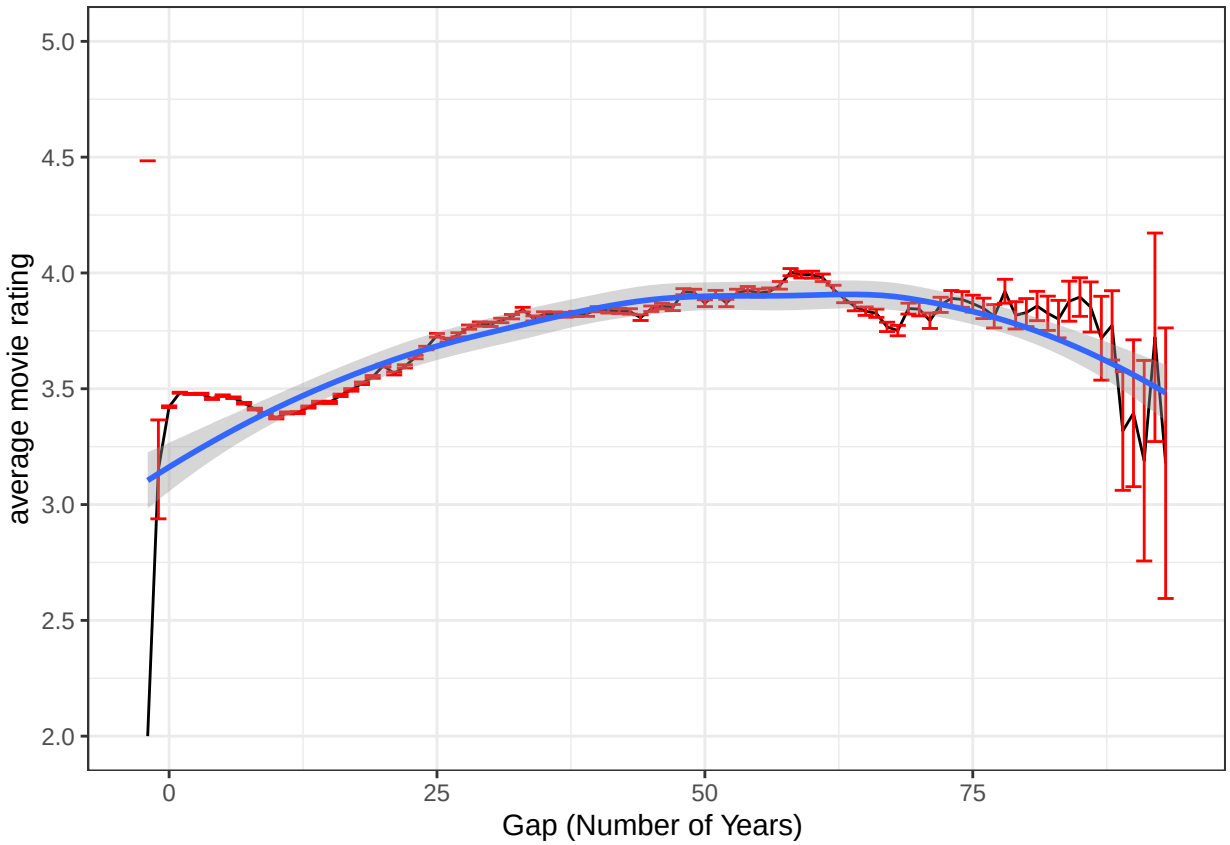


FIGURE 7: AVERAGE EFFECT BY GAP BETWEEN YEAR RATED AND YEAR MOVIE PREMIERED

- The gap between when a movie was rated and when it premiered ranged from -2 to 93 years. This indicates that there is some error in the data, as a user cannot rate a movie before it premiered. There were 161 ratings that fit this criteria. There are several ways to control for this error, one way is to delete these observations, another way to correct for this error without losing any observations is to recode all gap years that are negative to zero. This is assuming that the year the movie was rated was equal to the year it premiered. This is expected to have no material effect on the estimate.
- There is an upward trend (positive relationship) between the time a movie is rated and when it premiered up to 58 years. After the 70th year, this trend began to move downward (negative relationship). Suggesting the nostalgia effects (positive relationship), *things made in the past are better,* up to first 58 years and thereafter the technology effect (negative relationship), as the quality of the movie based on technology used may be too poor for persons to full enjoy the movie. However, it should be noted that the confidence intervals are largest in the outer years of the year gap, reflecting a relatively small number of ratings undertaken in those years. These represent movies that premiered in the earliest years (1915-1935).

Month rated

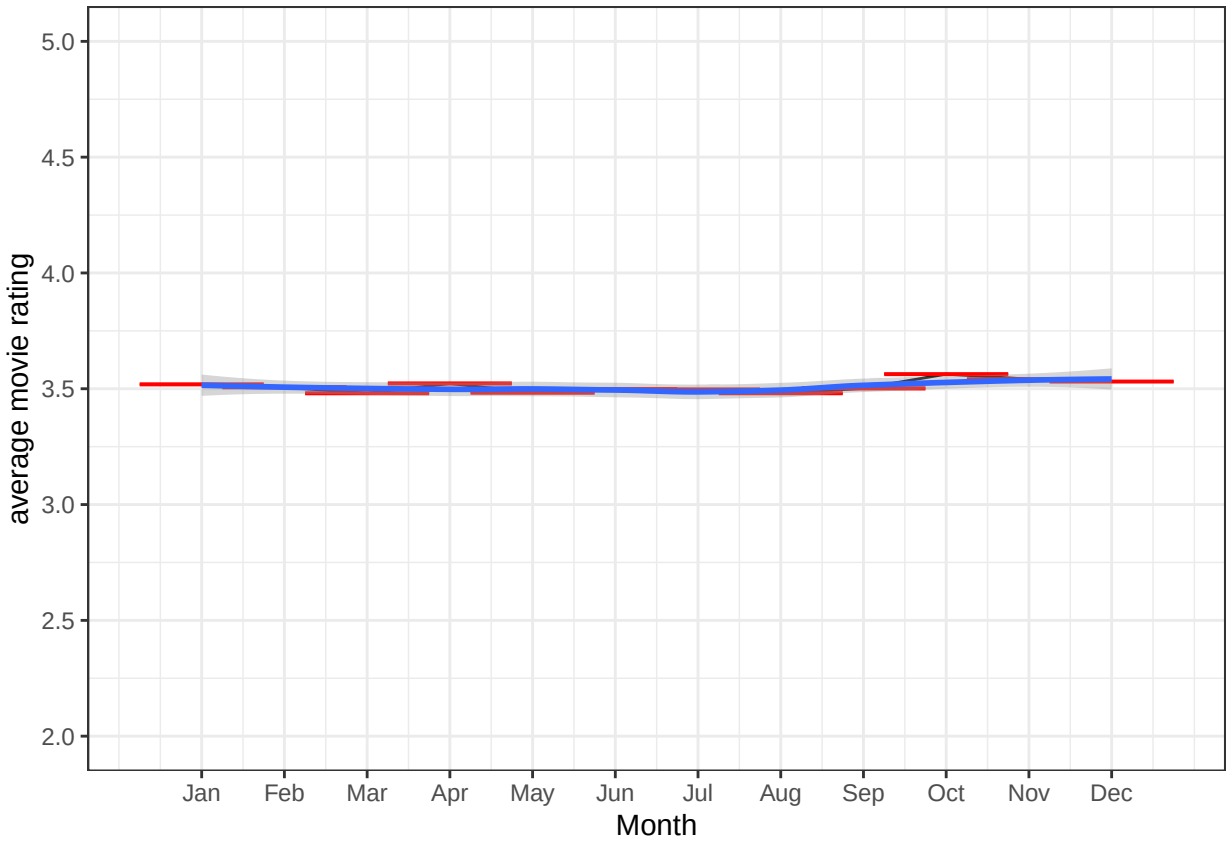


FIGURE 8: AVERAGE EFFECT BY MONTH RATED

- There is not much variability between the average rating when grouped by months (Figure 8). This suggests that this feature may not add significant value in distinguishing one movie from another.

Week rated

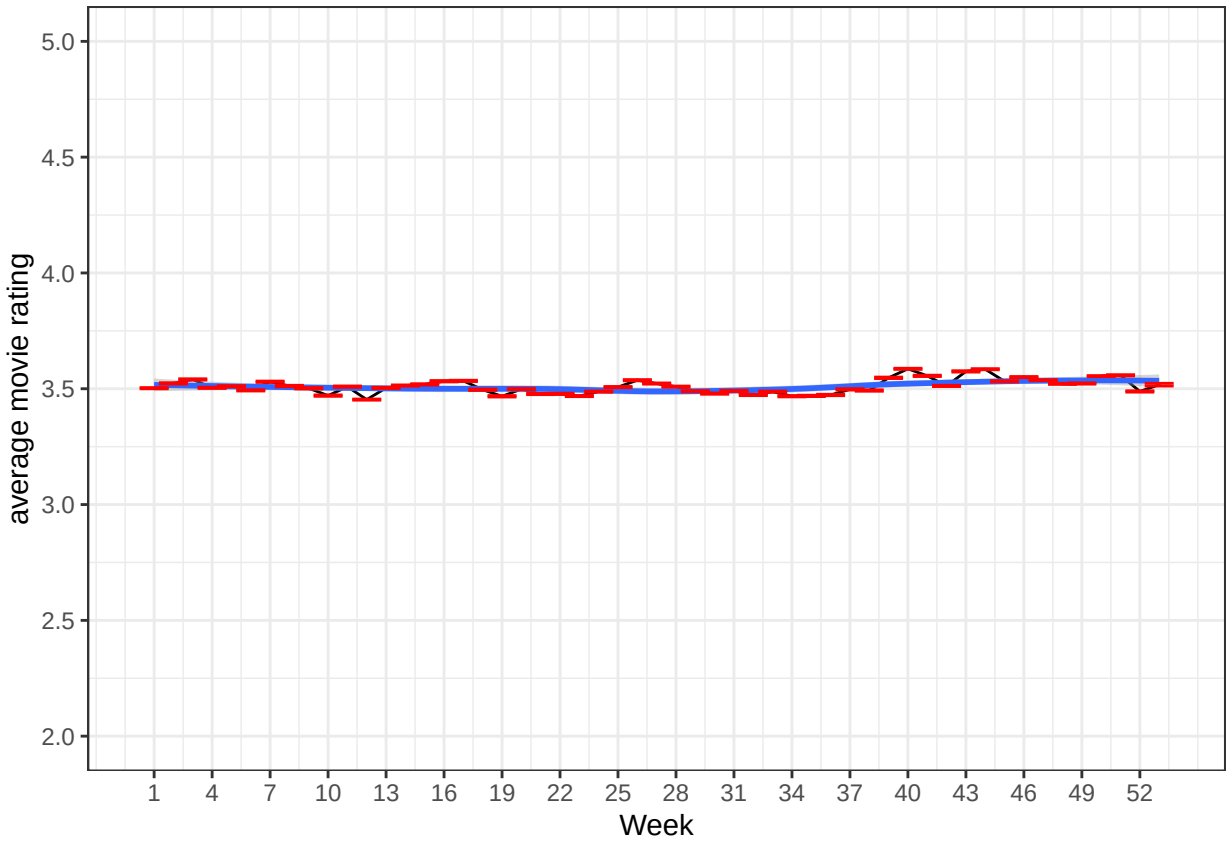


FIGURE 9 AVERAGE EFFECT BY WEEK RATED

- There is not much variability in the average when the data is grouped by weeks (Figure 9).

day rated

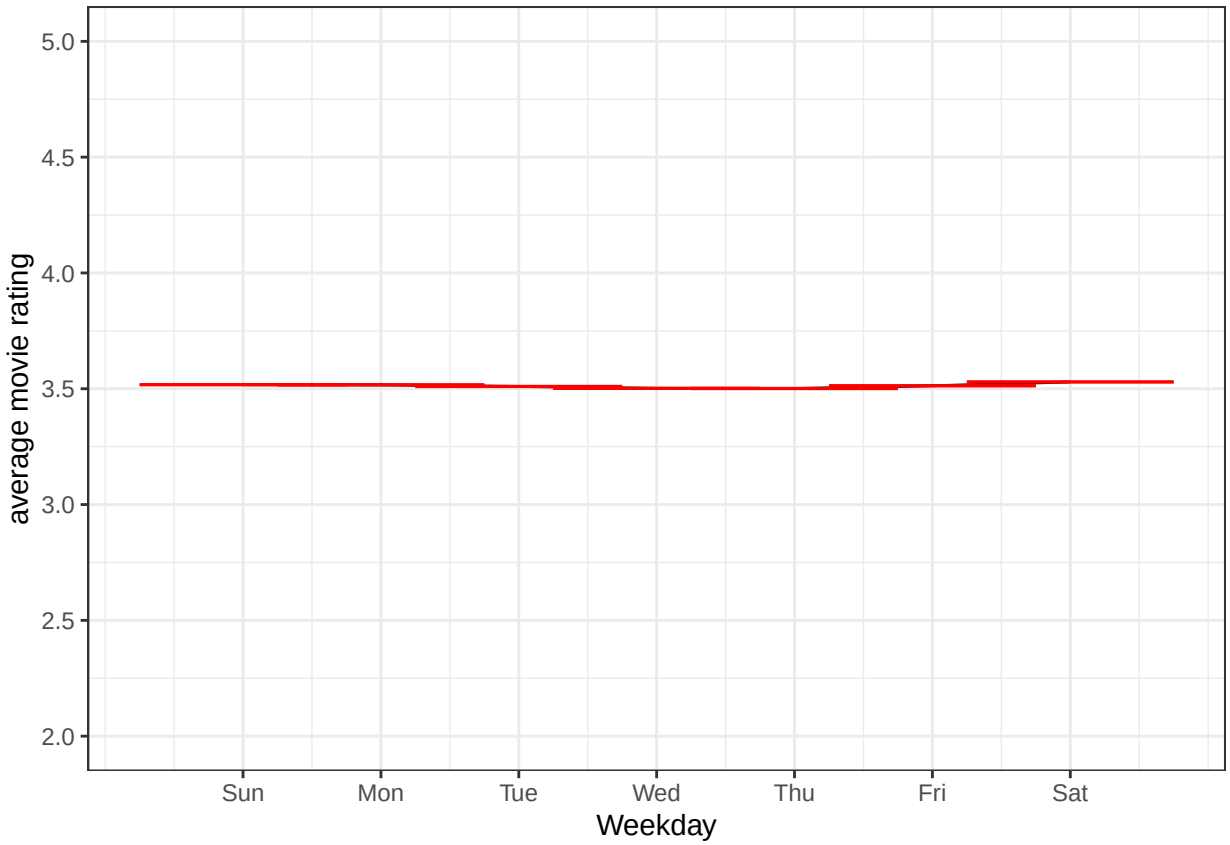


FIGURE 10: AVERAGE EFFECT BY WEEK DAY RATED

- There is not much variability when the data is grouped by days (Figure 10).

Table 10 presents a distributional breakdown of the date features by its mean, after adjusting for the discrepancies in the gap year features. That is, gap years that were negative were recoded to be equal to zero. The data suggests that the year features (premiere year, year rated and gap_year) have more variability and thus are potentially more useful in building a recommendation model than the month, week and day feature.

TABLE 10
DISTRIBUTION OF AVERAGE MOVIE RATINGS BY DATE FEATURES

Indicator	premiere year	year rated	year gap	month rated	week rated	day rated
min	3.270	3.426	3.179	3.481	3.453	3.501
Q1	3.510	3.467	3.555	3.494	3.488	3.506
median	3.749	3.505	3.815	3.504	3.509	3.513
Q3	3.902	3.562	3.857	3.525	3.532	3.517
max	4.055	5.000	4.003	3.563	3.586	3.529
Inter-Quartile Range	0.392	0.094	0.302	0.031	0.044	0.011
range	0.785	1.574	0.825	0.083	0.133	0.028

Table 11 presents a distributional breakdown of the data features by the number of times each of these features were rated. This is done to give insight in the reliability of some of the averages for each time period.

TABLE 11
DISTRIBUTION OF THE NUMBER OF MOVIE RATINGS BY DATE FEATURES

Indicator	premiere year	year rated	year gap	month rated	week rated	day rated

min	30	1	14	510,563	30,189	1,012,428
Q1	6,893	422,526	4,452	602,646	132,612	1,106,550
median	24,366	615,099	21,378	671,970	149,744	1,117,807
Q3	94,118	632,721	72,176	728,267	168,831	1,233,941
max	707,979	1,030,208	960,698	877,917	276,085	1,288,848
Inter-Quartile Range	87,225	210,195	67,724	125,621	36,219	127,391
range	707,949	1,030,207	960,684	367,354	245,896	276,420

SECTION III: METHODOLOGY

This section of the report provides the methodology that will be used to construct the movie recommendation model. The approach taken would incorporate lessons learned from the Netflix challenge (Bell and Koren 2007). Based on the exploratory data analysis, six predictors will be included in the model.

- Movie Effect (b_m), this will capture the average movie effect
- User effect (b_u), this will capture the average user effect while controlling for movie effects
- Genres effect (b_g), this will capture the average genre effect while controlling for movie and user effects
- Premiere Year effect (b_p), this will capture the average effect of the year the movie premiered while controlling for movie, user and genres effects
- Year Gap effect (b_{gap}), this will capture the average effect of the gap between when the movie was rated and when it premiered while controlling for movie, user, genre and premiere year effects
- Year Rated effect (b_r), this will capture the average effect of the year the movie was rated while controlling for movie, user, genres, premiere year and year gap effects

The goal is to estimate the rating of each movie by each user ($r_{m,u}$) using the six predictors. Our estimate of $r_{m,u}$ is $\hat{b}_{m,u}$. This suggests the following model:

$$\hat{b}_{m,u} = \bar{b} + \hat{b}_m + \hat{b}_u + \hat{b}_g + \hat{b}_p + \hat{b}_{gap} + \hat{b}_r$$

$$\hat{r}_{m,u} = \bar{b} + \hat{b}_m + \hat{b}_u + \hat{b}_g + \hat{b}_p + \hat{b}_{gap} + \hat{b}_r + \hat{e}_{m,u}$$

Where $\hat{e}_{m,u}$, represents the error term, the portion of $r_{m,u}$ that cannot be explained by the model. We will like to minimize the RMSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{m,u}\left(\hat{b}_{m,u} - r_{m,u}\right)^2}$$

One of the key findings of the exploratory data analysis is that there are small sample sizes for some of our estimates, for example, some movies are only rated once. This leads to overfitting of model parameters and an inflated variance. To control for that we will utilize a regularized model.

Regularized 1 (R1) will estimate each effect using a step wise approach using the edx_train dataset to build the model, and the edx_test dataset will be used to tune the penalty term. The penalty term that minimizes the RMSE will be utilized to build the model at each stage. Once an effect is estimated, it will be assumed to be fixed, and used in the estimate of the other effects. Therefore the penalty term used to estimate each parameter may be different.

Estimate of R1 effects:

$$Average - Effect : \bar{b}$$

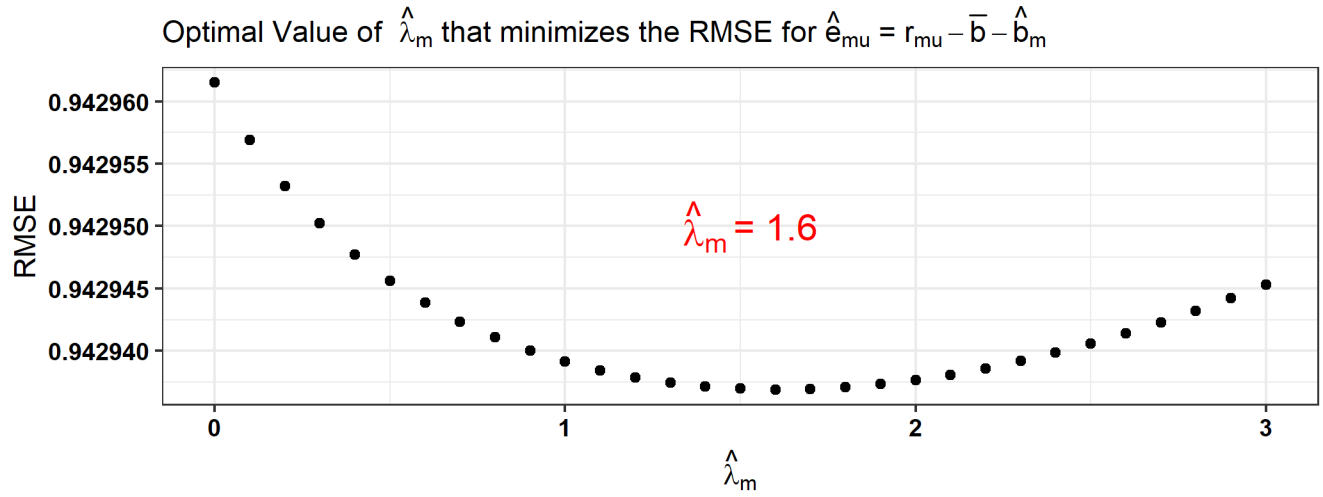
$$Movie - Effect : \hat{b}_m = \frac{1}{n_m + \lambda_m} \sum_{u=1}^{n_u} (r_{m,u} - \bar{b})$$

$$User - Effect : \hat{b}_u = \frac{1}{n_u + \lambda_u} \sum_{m=1}^{n_m} (r_{m,u} - \bar{b} - \hat{b}_m)$$

This is done for all the other effects, which will have a different value that optimizes lambda (penalty term) for each parameter.

SECTION IV:RESULTS

The model was developed using a step wise approach to assess the contribution of each additional variable. The penalty term was optimized using the lowest RMSE in the edx_test set for each variable (Figure 11 A, Figure 11B, Figure 11C).



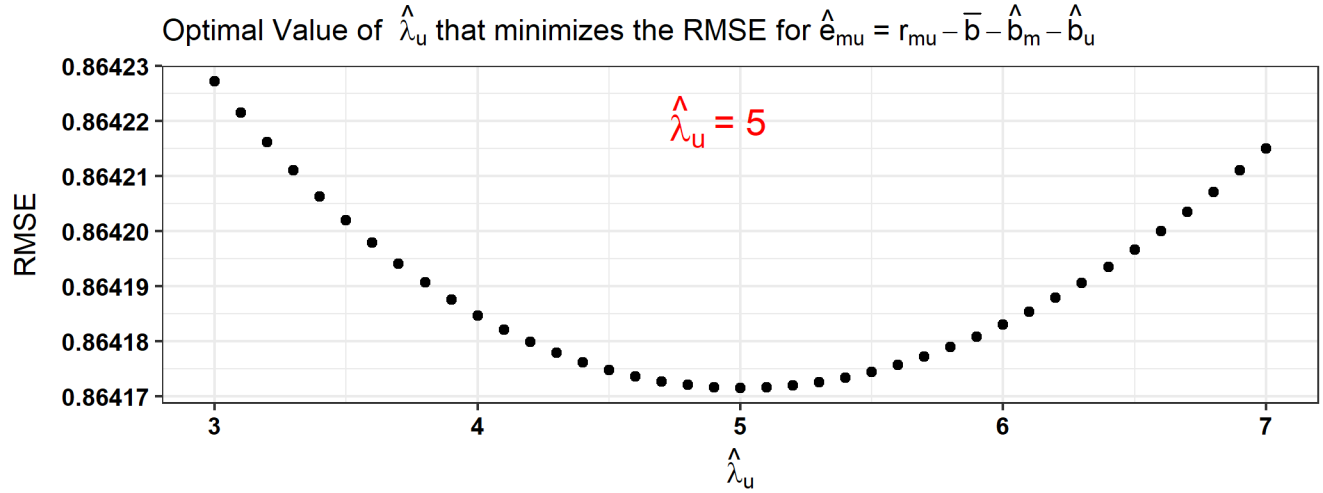


FIGURE 11A: Optimal value of λ_m and λ_u that optimize RMSE

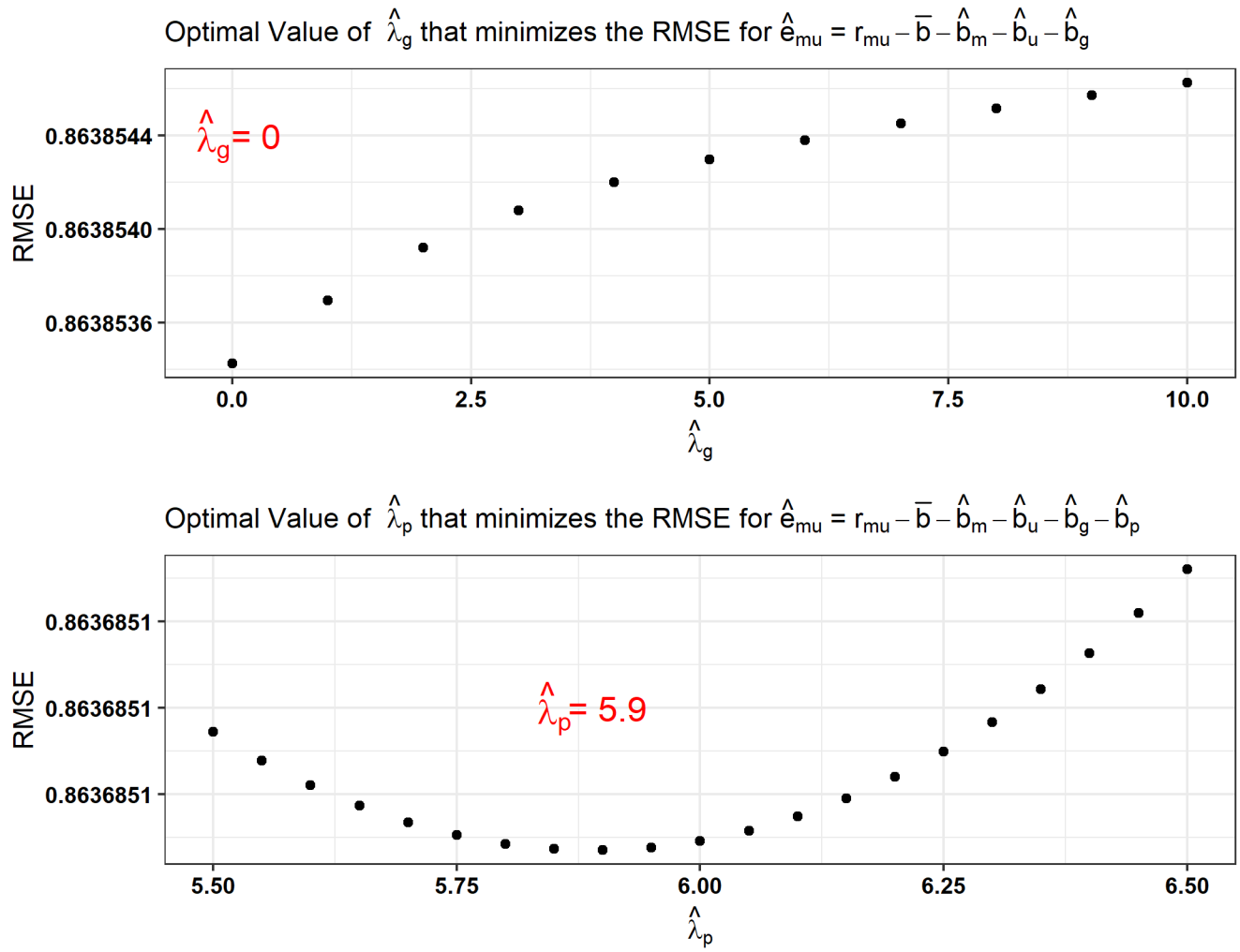


FIGURE 11B: Optimal value of λ_g and λ_p that optimize RMSE

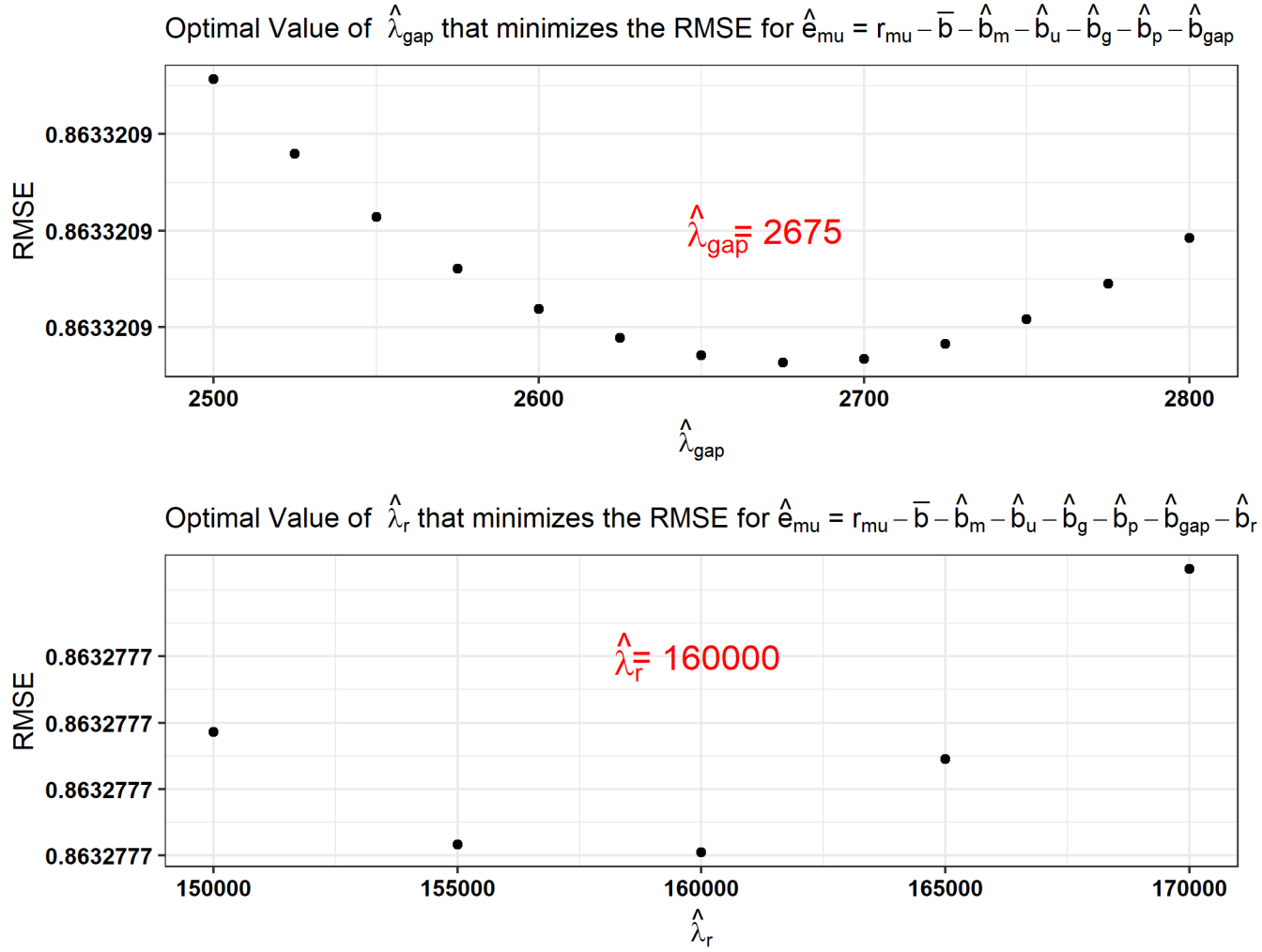


FIGURE 11C: Optimal value of λ_{gap} and λ_r that optimize RMSE

The greatest reduction in the RMSE came when the movieId and userId were included within the model (Table 12). The date effects contributed marginally to the reduction in the RMSE, when we controlled for the movie, user and genre effects.

TABLE 12
RMSE FOR EDX_TEST SET

METHOD	RMSE_R1
Average Effects	1.06005
Movie Effect	0.94294
Movie and User Effects	0.86417
Movie, User and Genres Effects	0.86385
Movie, User, Genres and Premiere Year Effects	0.86369
Movie, User, Genres, Premiere Year and Year Gap Effects	0.86332
Movie, User, Genres, Premiere Year, Year Gap and Year Rated Effects	0.86328

When the final model is applied to the validation dataset, an RMSE of 0.86433 was produced. It should be noted that without the date effects the RMSE would have been 0.86491, which is above the desired target of

0.86490.

SECTION V CONCLUSION

Summary

A model was developed to predict the rating a user will give to a movie, i.e. movie recommendation system. The model utilizes six predictors in model construction, which included the movie effect, user effect, genre effect and date effects. The date effects captures information on the year the movie premiered, the year the movie was rated, as well as the gap between when a movie was rated and when it premiered.

The model produced a RMSE of 0.86433, which was lower than the desired cutoff of 0.86490. The date effects were found to be important in achieving this goal, which highlights the importance of feature engineering in model development. Additionally, a regularization approach was also critical in model construct.

Limitations

- Due to lack of computer power (RAM capacity):
  - The parameters were approximation of the least square estimates.
  - Each lambda (penalty term) parameter was tuned individually instead of simultaneously, this would have reduced the potential accuracy of the model
  - Limitations were placed on feature engineering, which included generating interactive terms between the different predictors (e.g. date effects and genre - does the rating of genre horror, differs in October (Halloween) relatively to other months).
  - More sophisticated machine learning models were not used, such as Ridge, Lasso and Elastic Net.
- The model is only applicable to movies and users that are already part of the training set.

Further Extension

- Develop a model using a subset of the data and generate additional features and more sophisticated machine learning algorithms using cloud computing services (e.g. Amazon Web Services). Collect additional information on each movie as potential predictors, for example, the lead actress/actor and director of movie.

REFERENCES

Bell, Robert M., and Yehuda Koren. 2007. "Lessons from the Netflix Prize Challenge." *ACM SIGKDD Explorations Newsletter* 9 (2): 75–79. https://doi.org/10.1145/1345448.1345465.

Harper, F. Maxwell, and Joseph A. Konstan. 2016. "The Movielens Datasets." *ACM Transactions on Interactive Intelligent Systems* 5 (4): 1–19. https://doi.org/10.1145/2827872.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.

Kuhn, Max, and Kjell Johnson. 2019. *Feature Engineering and Selection*. Chapman; Hall/CRC. https://doi.org/10.1201/9781315108230.