MovieLense Project Report

Daniel Flores Orozco

2022-01-21

1. Introduction

MovieLens is a movie recommendation system that works as an online platform that provides personalized movie recommendations based on the user's and historical members' preferences. This platform collects different information, such as the user id, movie, rating, and when the rating was entered. The platform was created in 1997 by the GroupLens Research in the Department of Computer Science and Engineering at the University of Minnesota with the objective of collecting data on personalized recommendations. Since its creation, the MovieLens platform has generated tons of data. Currently, the MovieLense database is composed of approximately 11 million ratings of around 8,500 movies. The database includes the ratings on a scale of half (1/2) to 5 starts, movie information, and non-traceable user identity.

The different versions of the MovieLens dataset have been widely used during the last two decades in different applications, including personalization and recommendation research, the validation of novel data science techniques including summarization, data visualization, and machine learning algorithms. Moreover, the MovieLens databases are commonly used for education and academic purposes since it is publicly available and contains flexible data. In this document, the 10M MovieLens database structure was analyzed and visualized to ultimately use machine learning to progressively develop a model for predicting users' ratings based on historical ratings by users and movies, year of release, and movie genres.

2. Methods

2.1 MovieLense database

The 10M ratings dataset was downloaded from the MovieLens server (https://grouplens.org/datasets/movielens/10m/). Then, the database was adapted following the methodology recommended in the guidelines of the course to include six columns: *userId*, *movieId*, *rating*, *timestamp*, *title*, and *genres*. The *timestamp* column represented the time when the rating was provided, in seconds since midnight UTC January 1, 1970.

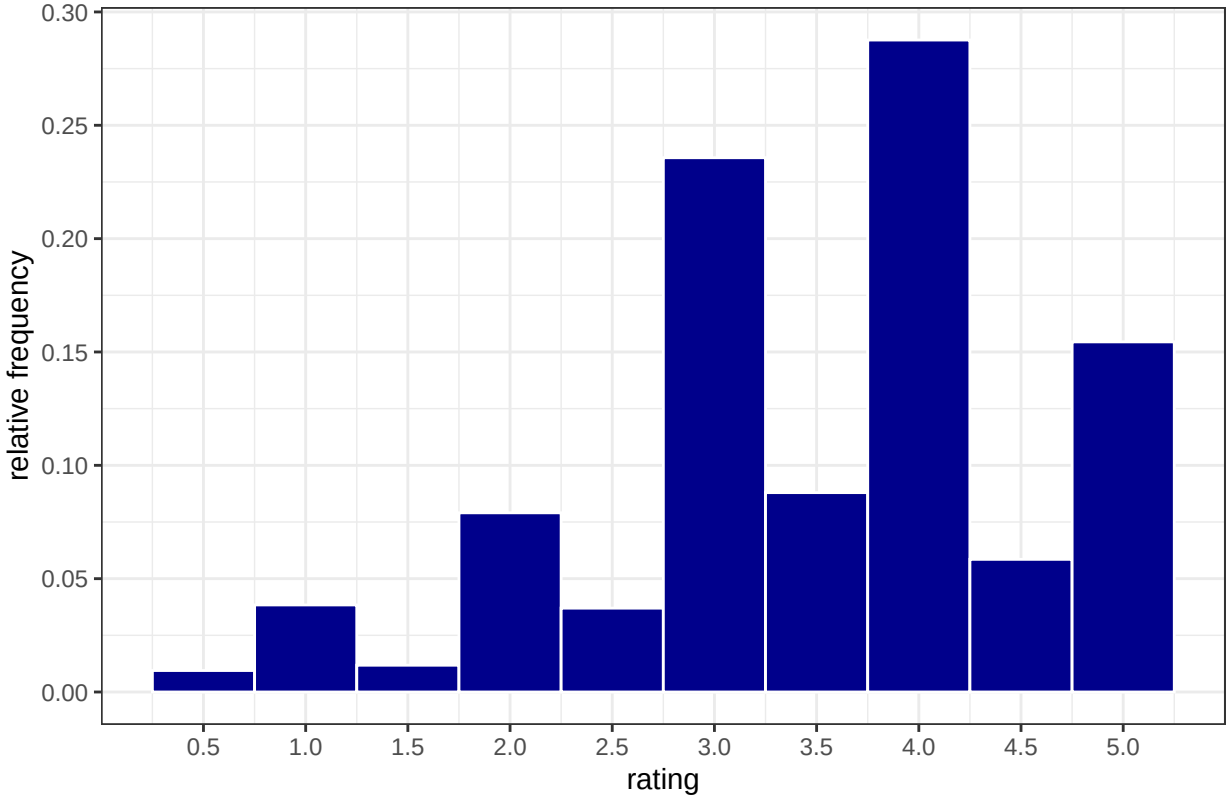
2.2 Working and validation data sets

The 10M data set was then split into validation and working data sets. The working dataset contained approximately 90% of the entries, while the other 10% was saved to evaluate the final model. Entries containing movies or users not included in the validation and working data sets were removed. Approximately 10% of the working data was set apart to test and evaluate the performance of the model.

2.3 Exploratory data analysis

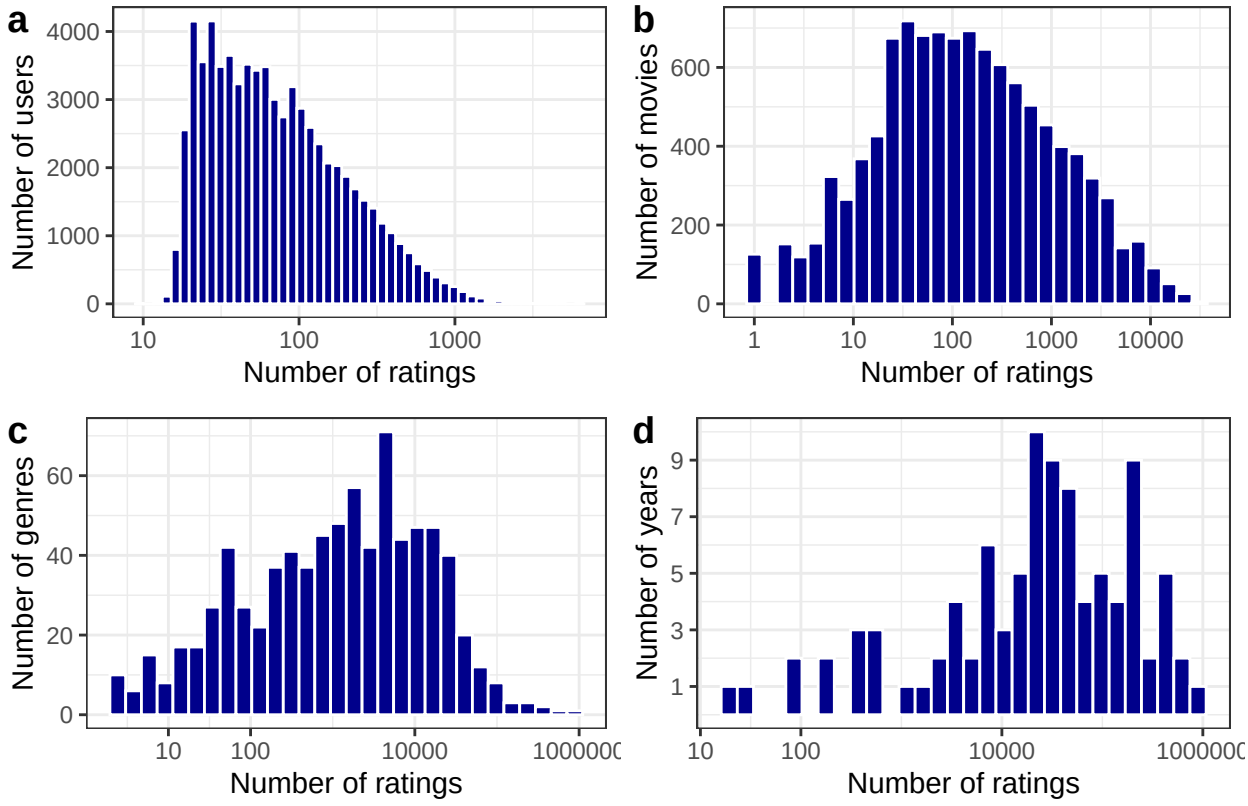
The first step was the analysis of the general features of the data. The working data set (edx) contained 9,000,055 entries. Each entry represented a rating given by a user to a specific movie. The data set contained ratings for 10,677 movies from 69,878 different users. There were more than 797 unique genres and 95 different years of movie release (1945-2008). The distribution of the ratings (Fig. 1) showed that round-number ratings were more common than half-number ratings. Also, ratings of 3 stars or greater were more common. The average rating of this data set was 3.51 starts with a standard deviation of 1.06.

Fig. 1 Distribution of the movie ratings



A deeper analysis of the data revealed more details about the data (Fig. 2). For instance, Fig. 2a showed that the distribution of ratings by users was right-skewed, indicating there were users with substantially more ratings than the average. The distribution of observations by movies (Fig. 2b) indicated that there were some movies with less than ten ratings and others with over 10,000. Similarly, the distribution of observations by genres (Fig. 2c) showed some genres with more than 10,000 ratings and others with less than 10. The distribution of the observations per year also showed high variability in the number of ratings.

Fig. 2 Distribution of observations



2.4 Loss function

The root-mean-square error (RMSE) was used to assess the accuracy of the different models. The RMSE is a commonly used loss function that measures the differences between values, usually the predictions generated by a model and the actual observed values. The RMSE (loss function) was defined as follows:

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

Where N is defined as the number of user,movie combinations, $y_{m,u}$ is the rating for the movie m by user u, and $\hat{y}_{m,u}$ the rating prediction for the movie m by user u. In this context, the RMSE was interpreted as the standard deviation, where a RMSE value of one represents an error in the prediction of one star. The goal of this project is to reduce the RMSE below 0.8649.

2.5 Models

The approach used in this project consisted of developing a series of models to predict ratings using as predictors the users, historical movie ratings, and genres and years of the movie release. Finally, the best model was used to predict ratings in the validation dataset.

2.5.1 Model 1: mean + error

The simplest model consisted in predicting the average of all movie ratings, assuming that the differences were caused by random variations. The model was defined as follows:

$$Y_{m,u} = \mu + \epsilon_{m,u}$$

Where μ represents the real rating for all movies and $\epsilon_{m,u}$ represents the sampling error from a distribution centered in zero. This model was fit to the train set derived from the edx data set and then tested in the test set also derived from the edx set. The μ of the train set was 3.51 with an SD of 1.06. The RMSE of this model is shown in Table 1.

2.5.2 Model 2: including all predictors

As shown in Fig. 3a, the distribution of ratings is not normally distributed. Similarly, the distribution of the model 1 residual is not normally distributed as seen in Figure 3. This indicated that $\epsilon_{m,u}$ could be influenced by other factors. The data (Fig. 2) suggested that userId, movieId, genres, and year of release might be causing some bias in the predictions. Therefore, in the second model, all the predictors (userId, movieId, genres, and year of release) were used to account for the variability of the predictions made with *Model 1* ($\epsilon_{m,u}$). The model was defined as follows:

$$Y_{m,u,g,y} = \hat{\mu} + b_m + b_u + b_g + b_y + \epsilon_{m,u}$$

Where b_m, b_u, b_g, and b_y represents the movie, user, genres, and year effects (bias), respectively. The individual effects were defined as:

$$b_m = \frac{1}{N}\sum_{m,u}(y_{m,u} - \hat{\mu})$$

Where $y_{m,u}$ is the rating of the movie m.

$$b_u = \frac{1}{N}\sum_{m,u}(y_{m,u} - \hat{\mu} - \hat{b}_m)$$

Where $\hat{b}_m$ is the movie effect.

$$b_g = \frac{1}{N}\sum_{m,u}(y_{m,u} - \hat{\mu} - \hat{b}_m - \hat{b}_u)$$

Where $\hat{b}_u$ is the user effect

$$b_y = \frac{1}{N}\sum_{m,u}(y_{m,u} - \hat{\mu} - \hat{b}_m - \hat{b}_u - \hat{b}_g)$$

Where $\hat{b}_g$ is the genres effect.

2.5.3 Model 3: regularized model

One of the main problems of *Model 2* is that it fails to provide good estimates to movies with small number of ratings For instance, the largest residuals (real - predicted ratings) came from movies with low observations (1-2 ratings), as seen in table 1. Based on this, the third approach consisted of using regularization to penalize large and noisy estimations derived from a small number of ratings.

Table 1: Larger Model 2 residuals

title	movieId	avg_rating	avg_res	obs
Suspended Animation (2001)	6933	0.5	3.776	1
My Flesh and Blood (2003)	7141	0.5	3.604	1
Quitting (Zuotian) (2001)	5579	5.0	3.470	1
Black Orchid, The (1958)	8876	0.5	3.211	1
Night Passage (1957)	6420	0.5	3.144	1
Two Friends (1986)	723	1.0	3.041	1
Only the Strong Survive - A Celebration of Soul (2002)	6340	0.5	2.986	1
Deathmaker, The (Der Totmacher) (1995)	6026	1.0	2.916	1
Gospel, The (2005)	39390	1.0	2.817	1
Dogs in Space (1987)	4101	1.0	2.759	1
Siam Sunset (1999)	6402	0.5	2.716	1
Eureka (1986)	6843	0.5	2.664	1
Grateful Dead Movie, The (1977)	30949	0.5	2.637	1
Everybody Wants to Be Italian (2007)	61467	4.0	2.559	1
Mutant Action (AcciÃ³n Mutante) (1993)	5822	4.5	2.451	1
Legal Deceit (1997)	1709	0.5	2.391	2
That Night in Varennes (La Nuit de Varennes) (1982)	6108	0.5	2.342	1
House on the Edge of the Park, The (La Casa Sperduta nel Parco) (1980)	6000	0.5	2.320	1
I'm Going Home (Je rentre Ã la maison) (2001)	6733	5.0	2.278	1
Next Step, The (1997)	1559	1.0	2.269	1

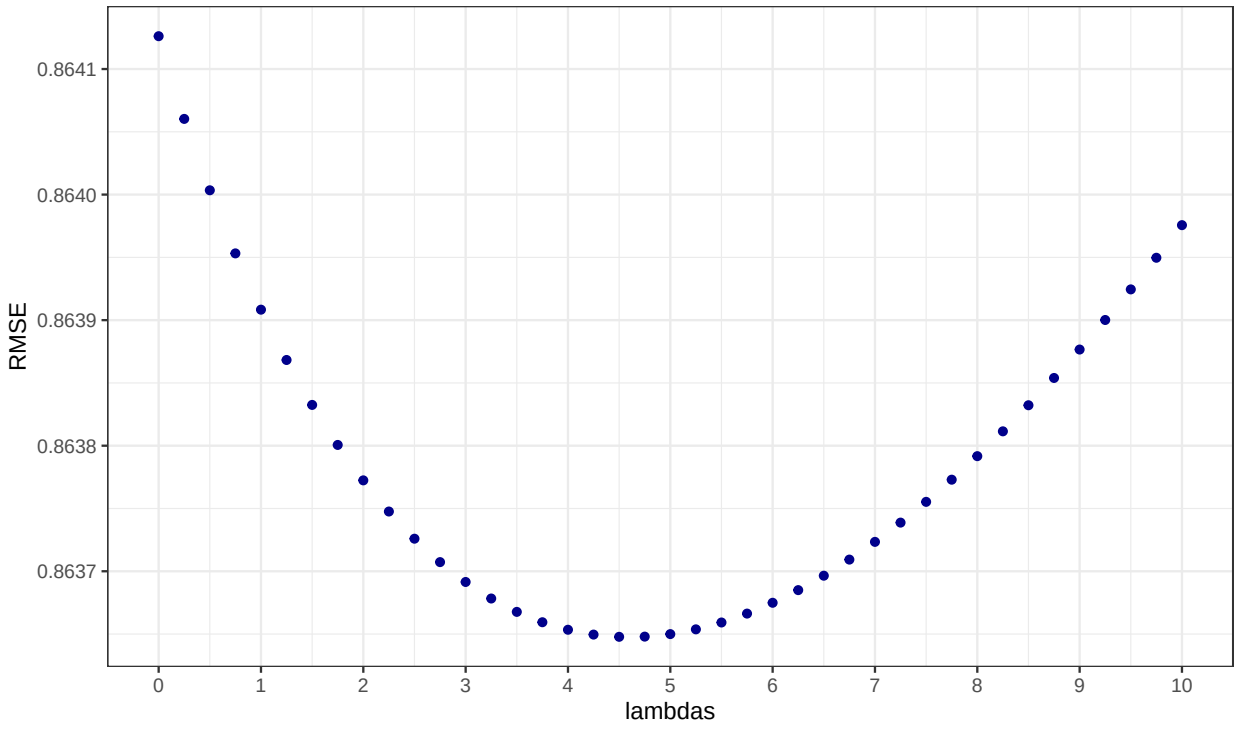
A similar equation than the one in the second model was used here, but with the predictors' effects ($\hat{b}$) adjusted with a penalty coefficient (λ) that minimizes the least-squares. Each predictor's effect was defined as follows:

$$\hat{b}_p(\lambda) = \frac{1}{\lambda + n_i} \sum_{u=1}^{n_i} (Y_{u,i,g,y} - \hat{\mu})$$

Where n_i is the number of ratings in each predictor group.

The optimal λ was found using cross-validation. Briefly, a series of λ values between 0-10 were used to find the value that minimizes the RMSE in the test set derived from the working data set (edx). The results of the tuning are shown in Fig. 4. The best λ was 4.5.

Fig. 3 Lambdas vs RMSE



2.6 Final test

The model with the best performance (*Model 3*) was fit to the validation data set, and the RMSE was evaluated. For this, the whole *edx* data set was used to develop the new *Model 3* version, increasing the number of observations, hence the statistical power.

3 Results

3.1 Models RMSE

The RMSE of the three models developed in this project are shown in Table 2. *Model 2*, which used *userId*, *movieId*, *genres*, and *year* of release as predictors, was able to reduce the RMSE to 0.8641 in the test set derived from the working data, which represents an improvement of over 18% compared to *Model 1*. *Model 3*, which included a regularization that penalized noisy estimates, also achieved a smaller RMSE (0.8636) compared to *Model 1* and *Model 2*, which represented an improvement of 0.6%.

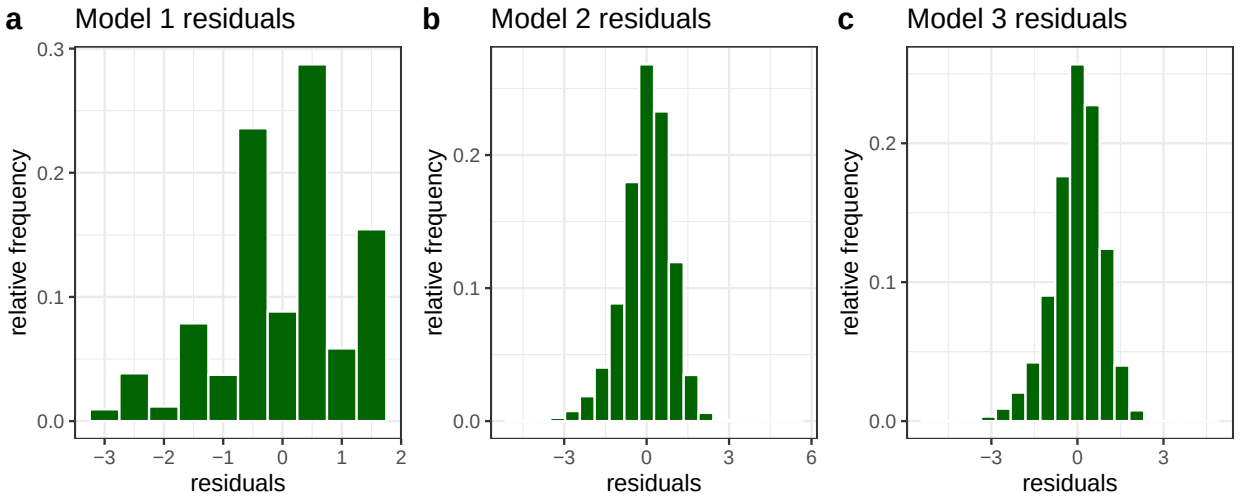
Table 2: Models RMSEs

Model	RMSE
Model 1: Just the mean	1.0601
Model 2: All predictors	0.8641
Model 3: Regularization	0.8636

The performance of the Models was also evaluated by analyzing the residuals (Fig. 4). As seen in Fig. 4a, the residual of *Model 1* were not normally distributed, which could be attributed to the effects of the

different predictors that were not considered. However, the *Model 2* and *Model 3* residual seemed to have normal distributions center in zero. In fact, the *Model 2* residuals mean was -0.001055, whereas the *Model 3* residuals mean was -0.001005, indicating an improvement in the estimates.

Fig. 4 Distribution of residuals



3.2 Final test

The final step in this project was to test the accuracy of the best performing model in the validation data set that was set apart since the beginning. The model was re-built using the whole working (edx) data set. The RMSE achieved was 0.8643 (Table 3), which was significantly below the target (0.8649).

Table 3: Final test in validation data set

Model	RMSE
Final Model 3: Regularized	0.8643

4 Conclusion

In this project, a machine learning approach was followed to build a model to predict movie ratings, also known as a movie recommendation system. The 10M MovieLens database was used as the source of data. The use of historical ratings grouped by users, movies, movie genres, and movie's year of release in a simple regression model with regularization to penalize noisy estimates proved to be efficient to reduce the RMSE below the target of 0.8649. The author is aware that the performance could be substantially improved by applying matrix factorization to the residuals. However, due to the limitations in computing power, this last step was not performed. Future projects are encouraged to implement matrix factorization to improve this recommendation system.