Movie Recommendation System

2020

Professional Certificate in Data Science

Contents

1 Introduction

1.1 Purpose

The purpose of this report is to develop a movie recommendation system through the use of machine learning algorithims. This report is a part of the Professional Certificate in Data Science program by HarvardX & edX.

The database used is a subset of the MovieLens dataset. It contains historical information about movie ratings, along with their associated movie titles & genres. The full dataset can be found here.

Before developing the model that will be used for recommendations, this report will firstly give a high-level overview of the sample data, as well the discuss key characteristsics of the sample data, and justify any trasnformations. We will then build multiple machine learning models to predict movie ratings and determine which of these models is the best at predicting ratings.

1.2 Executive Summary

After analysing the relationships between specific variables, the final prediction model built was a penalised least squares model, which takes into account the movie, user & genre effects of a rating. When compared to the validation dataset, the RMSE is 0.8632158.

2 Method & Analysis

2.1 Background

The Movielens sample data was paritioned into a 90-10 split. The 90% portion, which will be referred to as the "edx" dataset, will be the main focus of this report, as this is where will derive the training and test dataset for model building. The following sub-sections will provide an explaination into the structure of the edx data, as well as explain key transformations and insights. After this, we will then proceed into the model building stage. The 10% portion is the validation data, which will be used the the Results section of this report.

2.2 Initial Data Inspection

The edx dataset conatins 9000055 rows and 6 columns. There are 10677 different movies reviewed by 69878 unique users. The median number of reveiws for a single movie is 122, whilst the median number of movies reviewed by a single reviewer is 62. To give the reader a clear understanding of the edx dataset, the first few rows are shown below, as well as the summary statistics of each column:

Table 1: First few rows of the 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

Table 2: Summary Statistics of the edx dataset

userId	movieId	rating	timestamp	title	genres
Min. : 1	Min. : 1	Min. :0.500	Min. :7.897e+08	Length:9000055	Length:9000055
1st Qu.:18124	1st Qu.: 648	1st Qu.:3.000	1st Qu.:9.468e+08	Class :character	Class :character
Median :35738	Median : 1834	Median :4.000	Median :1.035e+09	Mode :character	Mode :character
Mean :35870	Mean : 4122	Mean :3.512	Mean :1.033e+09	NA	NA
3rd Qu.:53607	3rd Qu.: 3626	3rd Qu.:4.000	3rd Qu.:1.127e+09	NA	NA
Max. :71567	Max. :65133	Max. :5.000	Max. :1.231e+09	NA	NA

Since the purpose of this project is to develop a movie recommendation system based off ratings, below is a brief summary of the ratings within edx. Note that there are no 0 ratings.

Table 3: Ratings by count & proportion

Rating	Count	Proportion
4.0	2588430	29%
3.0	2121240	24%
5.0	1390114	15%
3.5	791624	9%
2.0	711422	8%
4.5	526736	6%
1.0	345679	4%
2.5	333010	4%
1.5	106426	1%
0.5	85374	1%

2.3 Data Cleaning

As shown in Table 1, the current form of the timestamp, title & genre columns are not ideal for analysis. This section describes the procedures made to the edx data to "clean" these columns. Non-numeric & non-integer columns were then removed from the edx dataset in order to make the size of the data smaller and to make the model building process more efficient. Duplicate rows were also removed. After cleaning, the edx dataset that will be used for model building now conatins 22 columns and 22rows.

2.3.1 Movie Title & Years

The title column in edx contains the name of the movie, as well as the year of release. Since the year of release could be a potential parameter in our model, it is ideal for us to separate this information into a new column. Note that the 'title' & "title_name" columns were not included in the training/test datasets for the model builing stage.

Table 4: First six rows of clean movie titles & year

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

title	title_name	year_made
Flintstones, The (1994)	Flintstones, The	1994

2.3.2 Rating Timestamps

The timestamps of user reviews in the edx data are unreadable to humans. As such, the first cleaning process with regards to the timestamp column is to convert the data into readable dates and times. Furthermore, additional columns were made to separate the date, weekday (where 1 = Monday, 2 = Tuesday, ..., 7 = Sunday), day, month, year and hour (24-hour format) of reviews, as these could be of potential value in the model building stage. Another column that was added was diff_years, which is the difference in the number of years between rating year and release year. Note that the "timestamp", "ts_date_time" & "ts_date" columns were not included in the training/test datasets for the model builing stage.

Table 5: First six rows of clean timestamps & diff_years

timestamp	ts_date_time	ts_date	ts_weekday	ts_day	ts_month	ts_year	ts_hour	diff_years
838985046	1996-08-02 11:24:06	1996-08-02	6	2	8	1996	11	4
838983525	1996-08-02 10:58:45	1996-08-02	6	2	8	1996	10	1
838983421	1996-08-02 10:57:01	1996-08-02	6	2	8	1996	10	1
838983392	1996-08-02 10:56:32	1996-08-02	6	2	8	1996	10	2
838983392	1996-08-02 10:56:32	1996-08-02	6	2	8	1996	10	2
838984474	1996-08-02 11:14:34	1996-08-02	6	2	8	1996	11	2

2.3.3 Genres

The genres column in edx concatenates the all the genres of a movie into one string. Should there be a need to include movie genres as a predictor in the movie recommendation model, this view of concatenating genres is not ideal for analysis. Thus, twenty new columns were added to the edx data, where each is a binary indicator of where a movie is a part of that genre. This allows us to separate each genre for analysis. Note that these all twenty extra columns were not used in the model builing stage.

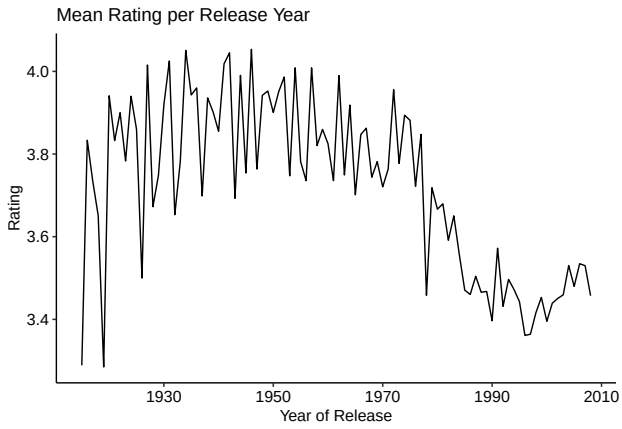
Table 6: First six rows of clean genres (7 of 20 shown)

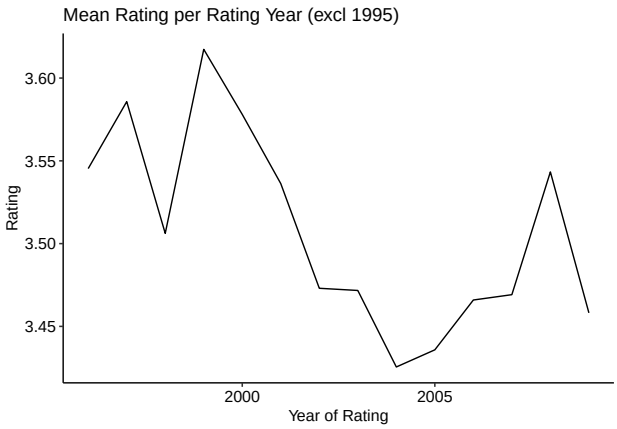
genres	Comedy	Romance	Action	Crime	Drama	Sci-Fi	Thriller
Comedy|Romance	1	1	0	0	0	0	0
Action|Crime|Thriller	0	0	1	1	0	0	1
Action|Drama|Sci-Fi|Thriller	0	0	1	0	1	1	1
Action|Adventure|Sci-Fi	0	0	1	0	0	1	0
Action|Adventure|Drama|Sci-Fi	0	0	1	0	1	1	0
Children|Comedy|Fantasy	1	0	0	0	0	0	0

2.4 Data Analysis

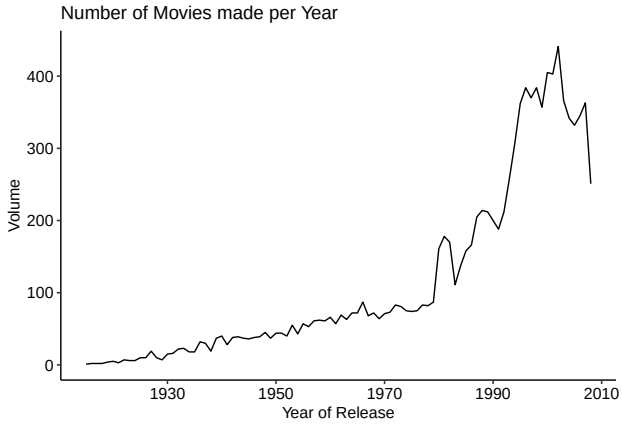
In this section, we will derive some preliminary insights from the cleaned edx data.

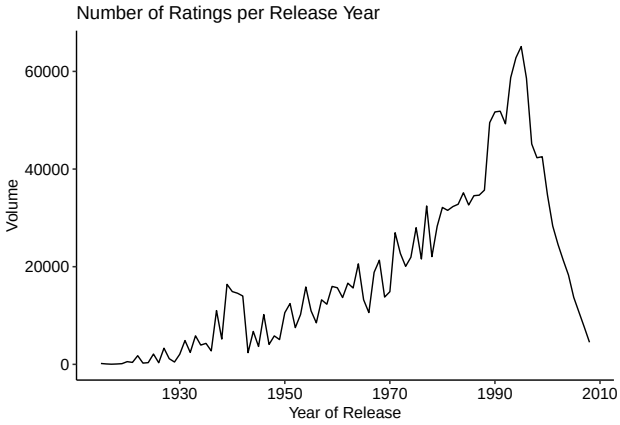
Firstly, we will look at the mean rating over time. For movie release years on the left, the mean rating has been declining since the 1970's. For rating years on the right, the mean rating has been declining. Note that ratings in 1995 were removed due to low volumes.



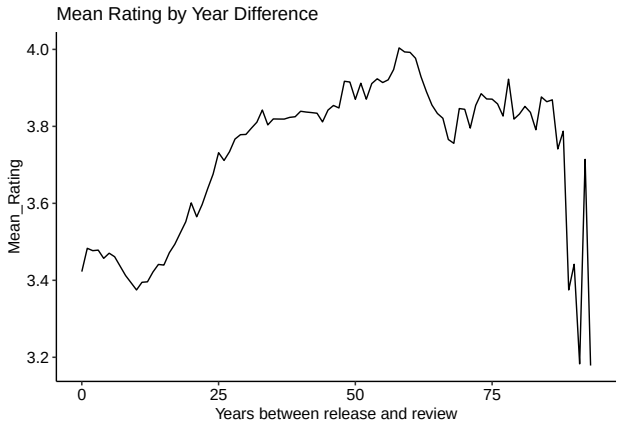


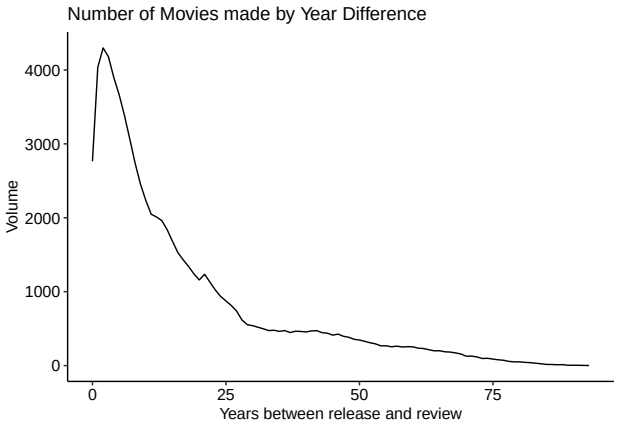
When charting the number of movies released per year, we can see that the volume grows signifiantly over time. This seems logical due to the general growth of the movie industry, increasing demand for movies, as well as the growth & accessibility of technology. Furthermore, the chart to the right shows the number of ratings per release year, which as expected, matches the trend in the number of movies made.





When looking ot the years between rating year and release year, we can see that whilst most movies are reviewed pretty quickly after relase, the mean rating tends to be higher for movies that are reviewed later on - this could potentially be due to nostalgia effects.





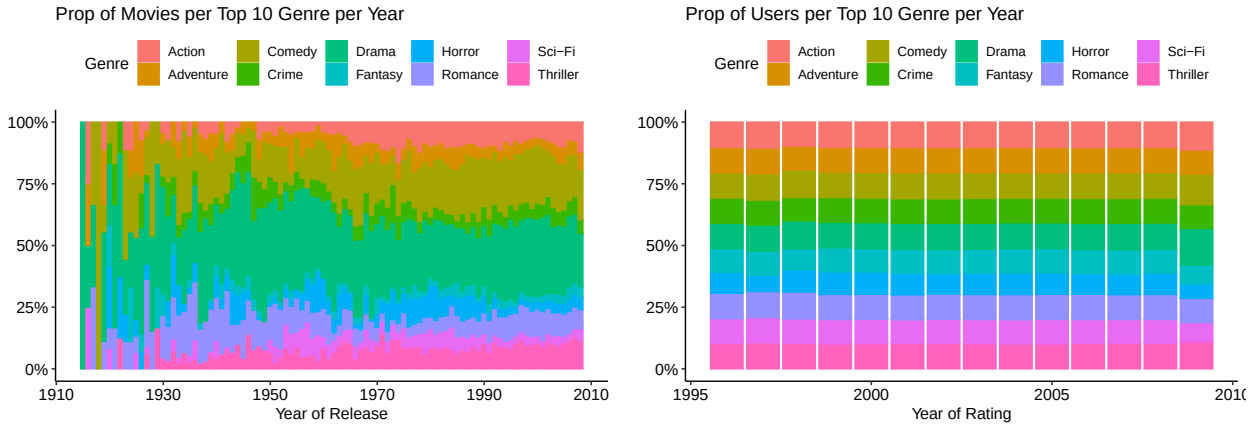
We will now look into trends per genre. Below is a summary table of the number of movies, number of user ratings & median rating, per genre. The proportions of each value against all other genres are also recorded, and the table is ordered by movie volume. We can see that the median rate per genre is between

3.26 (Horror) and 3.78 (Drama). One interesting finding to note is that the number of users per genre is fairly constant, despite the disparity of movie volumes per genre.

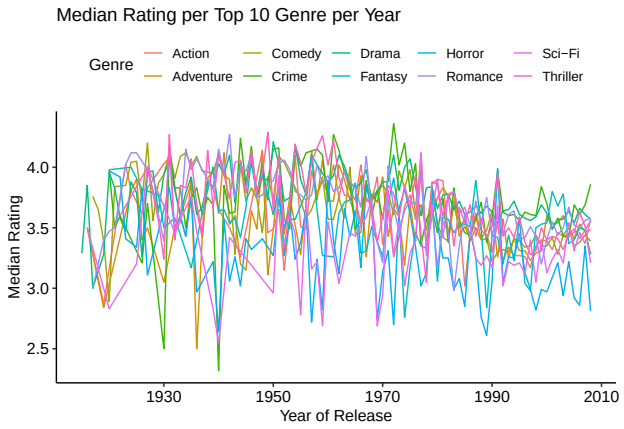
Table 7: Number of movies, User reviews & Median rating, per Genre

Genre	Movies_Count	Movies_Prop	Users_Count	Users_Prop	Ratings_Median	Ratings_Prop
Drama	5336	25%	93245	7%	3.780	5%
Comedy	3703	17%	92452	7%	3.690	5%
Thriller	1705	8%	91040	6%	3.675	5%
Romance	1685	8%	90602	6%	3.705	5%
Action	1473	7%	90839	6%	3.565	5%
Crime	1117	5%	88859	6%	3.700	5%
Adventure	1025	5%	90441	6%	3.535	5%
Horror	1013	5%	75309	5%	3.260	5%
Sci-Fi	754	3%	87160	6%	3.395	5%
Fantasy	543	3%	85644	6%	3.540	5%
Children	528	2%	80072	6%	3.480	5%
War	510	2%	80807	6%	3.760	5%
Mystery	509	2%	79045	6%	3.710	5%
Documentary	481	2%	32125	2%	3.710	5%
Musical	436	2%	72513	5%	3.570	5%
Animation	286	1%	73505	5%	3.570	5%
Western	275	1%	56378	4%	3.550	5%
Film-Noir	148	1%	38746	3%	3.895	5%
IMAX	29	0%	7118	1%	3.400	5%
(no genres listed)	1	0%	7	0%	3.640	5%

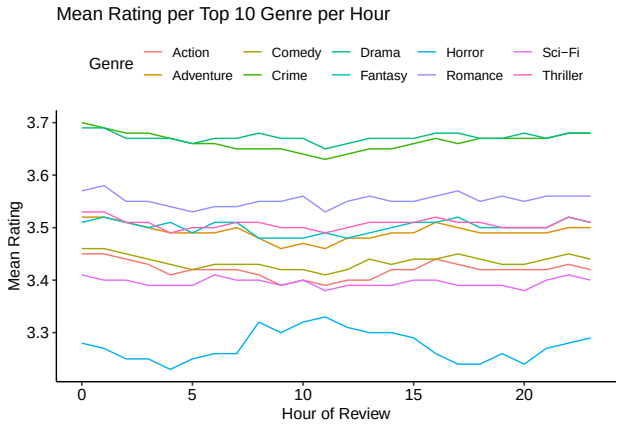
The next few charts will now look at genres over time. Note that we have only considered the top ten genres by movie volume for plotting, as this comprises 86% of the entire edx dataset. We will firstly examine the volume proportions of movies per release year by genre. We can see that since the 1930's, the proportion of genres has remain faily stable over time, with Drama and Comedy being the most popular genres made. When charting the volume propotions of user ratings over time by genre, we can see that each proportion is stable over time.



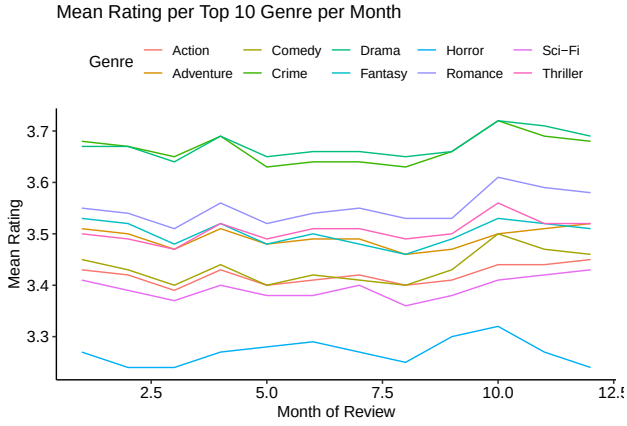
The below chart shows the median rating over time by genre. Whilst being rather volatile, this shows that all genres have had a median rating generally between 3 and 4 over all release years.



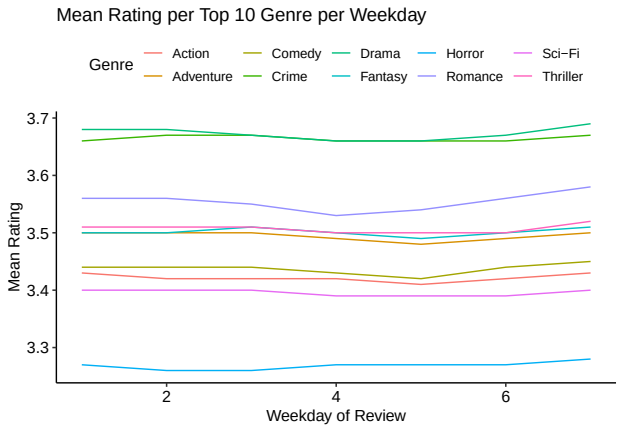
Since we have separated the timestamp column to different time values, we will now look into any potential trends in the mean rating for each genre per hour, month and weekday of review. From the chart below, we can see that the hour of review does not tend to have any significant impact on ratings per genre. However, the mean ratings of each genre are significantly separated.



Per month, the mean rating reamains fairly stable throughout the year. But again, the genre trends are clearly separated.



Finally, we will now look into the trend in the weekdays of reviews. Once again, the rating does not appear to signifiancantly alter over the week, but genres are separted.



2.5 Model Building

Noting the analysis that we done in previous sections of this report, in this section we will now look into bulding a few machine learning models to predict the ratings of movies. To produce the training & testing data, the cleaned edx dataset was split 80/20. Also, with regards to the "genres" column, each row was separated into its specific genre, as this will be used later on in the model building process. There are 18696902 rows in the training data and 4674203 rows in the testing data. Models were build using the training set, and then compared to the test set via RMSE, in order to determine if additional parameters are need before a final model can be selected. The final model, as well as the use of the validation set, are explained in the next section of this report.

2.5.1 Mean

The first model that will be examined is simple mean of all ratings from the test dataset. This is shown in the formula below, where u, m refers to each user & movie pairing.

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

Whilst this is not expected to be the final model, it does provide us with a benchmark. When compared to the test dataset, the RMSE of this model is 1.051687.

2.5.2 Movie Effects

To improve on the above model, the average rating of each movie should be added to the model, as this will remove any movie bias from our model. When compared to the test dataset, the RMSE of this model is 0.9406938.

$$rating_{u,m} = \mu + movie_m + \epsilon_{u,m}$$

2.5.3 User Effects

In addition, some users will tend to follow a certian patten of ratings. Thus, any user bias should be removed by taking the mean rating for each user as well. When compared to the test dataset, the RMSE of this model is 0.8573044.

$$rating_{u,m} = \mu + movie_m + user_u + \epsilon_{u,m}$$

2.5.4 Genre Effects

As seen in the previous Analysis section, there appears to be biases in ratings based on Genre, where ratings by genres tended to remain fairly stable over all different time measures. As such, these biases should be accounted for in our model, as shown below. As previously mentioned, the "genre" column had concatenations of all the genres. When compared to the test dataset, the RMSE of this model is 0.8572942.

$$rating_{u,m} = \mu + movie_m + user_u + genre_m + \epsilon_{u,m}$$

2.5.5 Penalised Least Squares

Movies with ratings that are signifiantcly above or below our estimate will constrian our model's predictability. By accounting for the sample size of each effect through the use of a penalisation parameter, λ, we are able to make the model less variable. The penslisation parameter is applied to each mean grouping in the following way:

$$pen.movie_m = \sum(rating_{u,m} - \mu)/(n + \lambda)$$

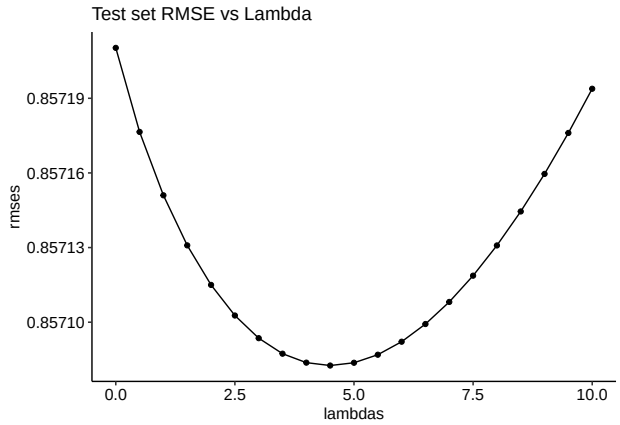
$$pen.user_u = \sum(rating_{u,m} - \mu - pen.movie_m)/(n + \lambda)$$

$$pen.genre_m = \sum(rating_{u,m} - \mu - pen.movie_m - pen.user_u)/(n + \lambda)$$

Resulting in a similar model as previously described, but with different parameters:

$$rating_{u,m} = \mu + pen.movie_m + pen.user_u + pen.genre_m + \epsilon_{u,m}$$

To pick a suitable value for the penalisation parameter, the previous model with movie, user & genre effects accounted for was used to pick the value of λ that returns the lowest RSME against the test set. The results of each run are shown in the chart below, which shows us that the most suitable value is $\lambda = 4.5$.



When compared to the test dataset, the RMSE of this model is 0.8570826.

3 Results

The final model that will used is the penalised least squares model, which accounts for movie, user and genre effects in ratings. This was selected as it had the lowest RMSE against the test dataset.

Table 8: Model Results against Test Set

Model Type	Test Data RMSE (to 5 d.p.)
Mean	1.05169
Mean + Movie	0.94069
Mean + Movie + User	0.8573
Mean + Movie + User + Genre	0.85729
Penalised Mean + Movie + User + Genre	0.85708

To check the accuary of our final model, the RSME of predicted ratings were calculated against the ratings in the validation dataset, which results in a final RMSE of 0.8632158.

4 Conclusion

4.1 Summary of Findings

As shown by the final model's RSME against the validation dataset, this model is a prettygood indicator of movie ratings from the sample Movelens data. The analysis showed that the movie itself, the specifc user, as well as the genre of a movie, are the primary indicators of a movie's rating.

4.2 Suggested Improvements

Whilst the above 3 effects were modeled, it should noted that there could potentially be more parameters that we could consider in the model. For example, the date & timestamp of a rating could have been used to predict ratings. This was not considered in the model reported as the validation RMSE was already sufficiently sound once the Genre effect was condisered. Additonally, whilst it is known that a model with more parameters is always going to lead to a better fit of the data (i.e. a lower RMSE), there must be a balance between model fit and predictability. The addition of too many parameters would lead to over-fitting and lower predictive power of a model.

Other suggestion for future models would be to consider other machine learning techniques such as K-Nearest Neghours or XGBoost. Due to machine limitations, I was unable to run a complex machine learning model with such a large dataset as the edx data. However, if the dataset was smaller, it would be more feasible to try out other models.