Movie Recommendation System Capstone Project

Fahad Bin Imtiaz

2024-10-10

1. Introduction / Executive Summary

The goal of this project is to develop a movie recommendation system based on the MovieLens 10M data set. The data set contains over *10 million ratings* from more than *69,000 users* on *10,000 movies.* The primary objective of this project is to predict how users will rate movies they have not yet seen, by developing a movie recommendation model system. To evaluate the performance of the model, we will use **Root Mean Squared Error (RMSE)** as the primary metric as explained in the Capstone project, with a goal of minimizing the prediction error using RMSE (where the lower the RMSE, the better the recommendation system). The target is to build a model with an RMSE of **less than 0.8649**.

The following steps were undertaken to complete this project: 1. Data cleaning and pre-processing. 2. Exploratory data analysis and visualizations. 3. Model building, starting with simple models and progressing to regularized models. 4. Final evaluation using a holdout test set.

First of all, we will create an edx set and final_holdout_test set from the code provided by the course.

We also need Scales package for proper sclaing of graphs and plots for data visualization.

2. Methods / Analysis

We have already set up the data set with the edx and final_holdout_test splits provided by the course in the previous code above. The next steps will focus on understanding the data. We'll explore user-movie interactions, rating distributions, and genre distributions to get insights about our data set.

Step 2.1 Dataset and Preprocessing

We need to confirm the structure and contents of the edx data set before proceeding with modeling, because this is crucial for understanding how users rate movies and for determining any potential data pre-processing needs. From the code provided by the course, we will only use training data set (edx) to train our model and implement final model on our validation set (final_holdout_test). We have observed that in this data set, we have *69878* unique users and *10677* unique movies. We also have checked the structure of the data set for initial analysis, including if data set contains any missing values, and none were found. The genres field, containing multiple genres for each movie, is required to be distributed into individual genre for further insights.

```
## 'data.frame':    9000055 obs. of  6 variables:
##  $ userId   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId  : int  122 185 292 316 329 355 356 362 364 370 ...
##  $ rating   : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ timestamp: int  838985046 838983525 838983421 838983392 838983392 838984474 838983653 838984885 83
##  $ title    : chr  "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
##  $ genres   : chr  "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|A
```

```
##   userId movieId rating timestamp                         title
## 1      1     122      5 838985046              Boomerang (1992)
## 2      1     185      5 838983525               Net, The (1995)
## 4      1     292      5 838983421               Outbreak (1995)
## 5      1     316      5 838983392               Stargate (1994)
## 6      1     329      5 838983392 Star Trek: Generations (1994)
## 7      1     355      5 838984474       Flintstones, The (1994)
##                          genres
## 1                Comedy|Romance
## 2         Action|Crime|Thriller
## 4  Action|Drama|Sci-Fi|Thriller
## 5       Action|Adventure|Sci-Fi
## 6 Action|Adventure|Drama|Sci-Fi
## 7       Children|Comedy|Fantasy
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.500   3.000   4.000   3.512   4.000   5.000
```

```
## Total missing values in edx dataset:  0
```

```
## Number of unique users:  69878
```

```
## Number of unique movies:  10677
```

Step 2.2 Data Wrangling / Exploratory Data Analysis

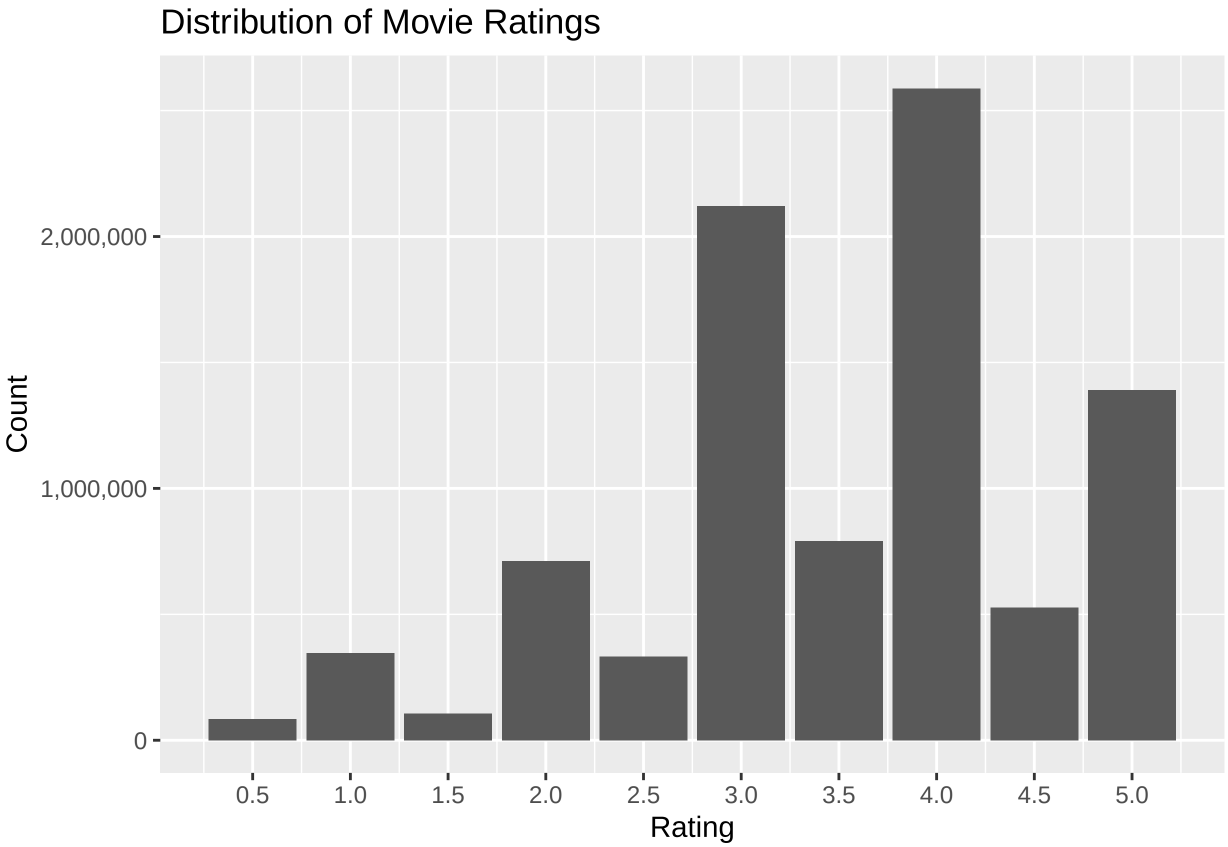
We require a basic exploration of the data set to better understand the distribution of ratings and user behavior in data visualization. *Pulp Fiction (1994)* has got the highest rating count of *31362*. Top 10 users have given more then 3100 movie ratings. It was observed that half star ratings were less then full star ratings and Rating **4** was given with a count of *2588430*. During data exploration, we observed that the *genres* are pipe-separated values. Therefore, genres were separated which gave different aspect of data set. It was necessary to extract them for more consistency, robust and precise estimate. It was observed that **Drama**, **Comedy**, and **Action** movies being the most frequently rated.

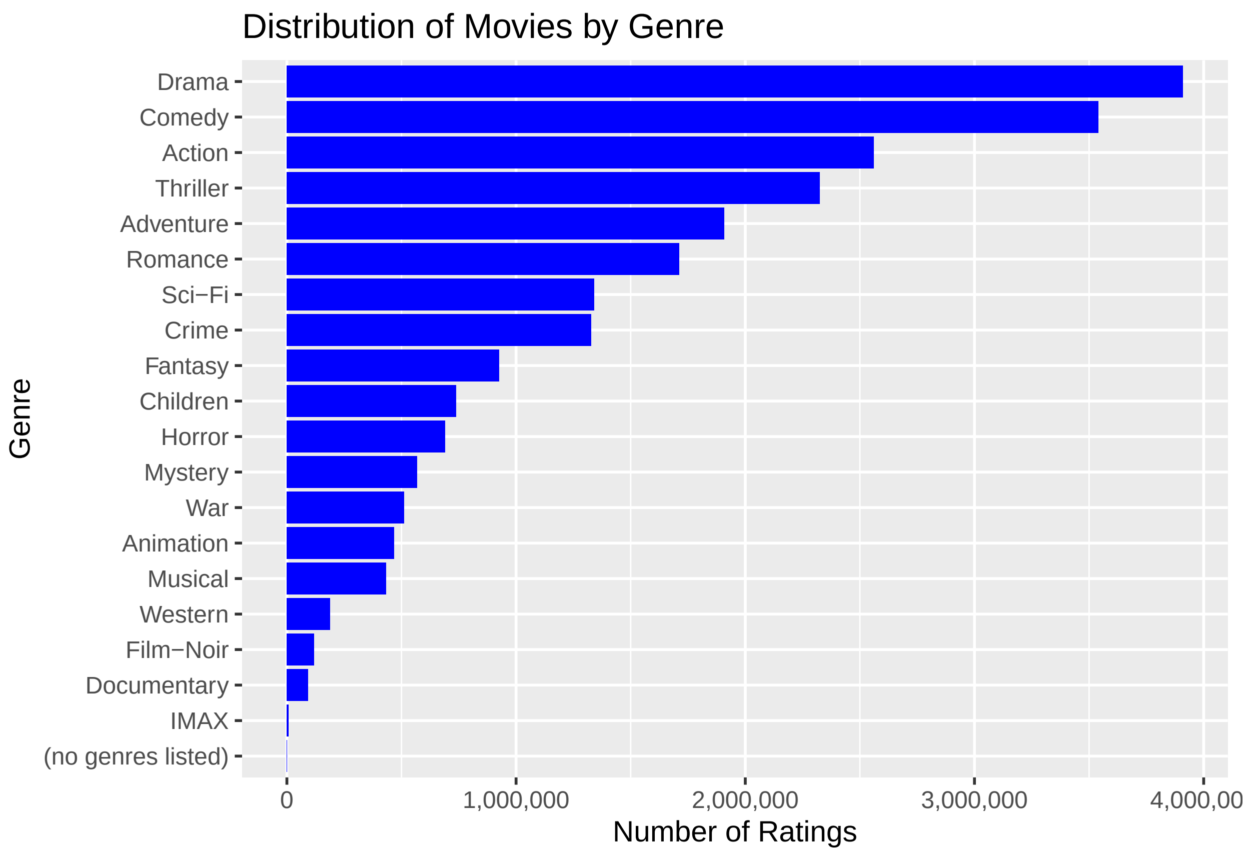
```
## Top Ten Most Rated Movies are:
```

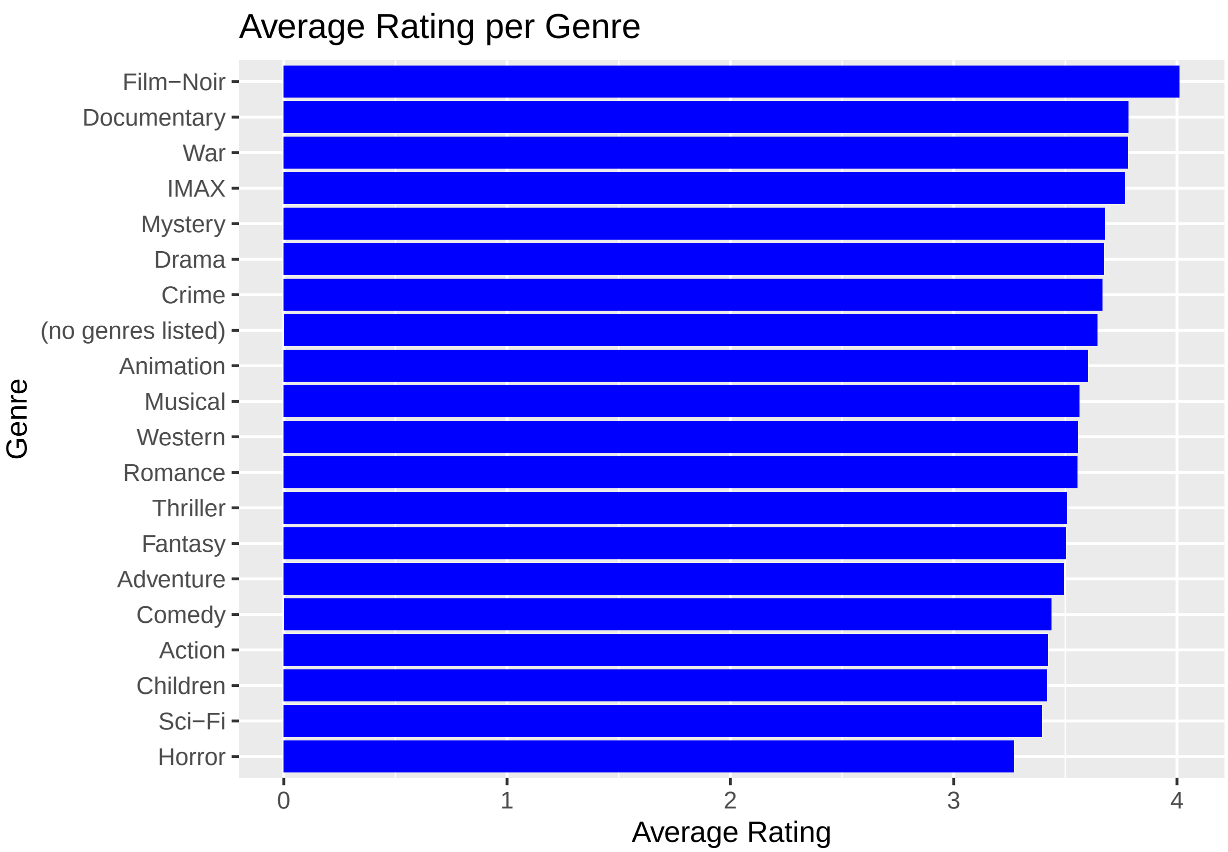
```
## # A tibble: 10 x 2
##    title                                                        rating_count
##    <chr>                                                               <int>
##  1 Pulp Fiction (1994)                                                 31362
##  2 Forrest Gump (1994)                                                 31079
##  3 Silence of the Lambs, The (1991)                                    30382
##  4 Jurassic Park (1993)                                                29360
##  5 Shawshank Redemption, The (1994)                                    28015
##  6 Braveheart (1995)                                                   26212
##  7 Fugitive, The (1993)                                                25998
##  8 Terminator 2: Judgment Day (1991)                                   25984
##  9 Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)        25672
## 10 Apollo 13 (1995)                                                    24284
## Top Five Ratings are:

## # A tibble: 5 x 2
##   rating   count
##    <dbl>   <int>
## 1    4   2588430
## 2    3   2121240
## 3    5   1390114
## 4    3.5  791624
## 5    2    711422

## Half-star ratings are less common than whole-star ratings:  TRUE

## Top Ten Active users are:

## # A tibble: 10 x 2
##    userId rating_count
##     <int>        <int>
##  1  59269         6616
##  2  67385         6360
##  3  14463         4648
##  4  68259         4036
##  5  27468         4023
##  6  19635         3771
##  7   3817         3733
##  8  63134         3371
##  9  58357         3361
## 10  27584         3142

## Genre Counts are:

## # A tibble: 20 x 2
##    genres               count
##    <chr>                <int>
##  1 Drama              3910127
##  2 Comedy             3540930
##  3 Action             2560545
##  4 Thriller           2325899
##  5 Adventure          1908892
##  6 Romance            1712100
##  7 Sci-Fi             1341183
##  8 Crime              1327715
##  9 Fantasy             925637
## 10 Children            737994
## 11 Horror              691485
## 12 Mystery             568332
## 13 War                 511147
## 14 Animation           467168
## 15 Musical             433080
## 16 Western             189394
## 17 Film-Noir           118541
## 18 Documentary          93066
## 19 IMAX                  8181
## 20 (no genres listed)       7
```

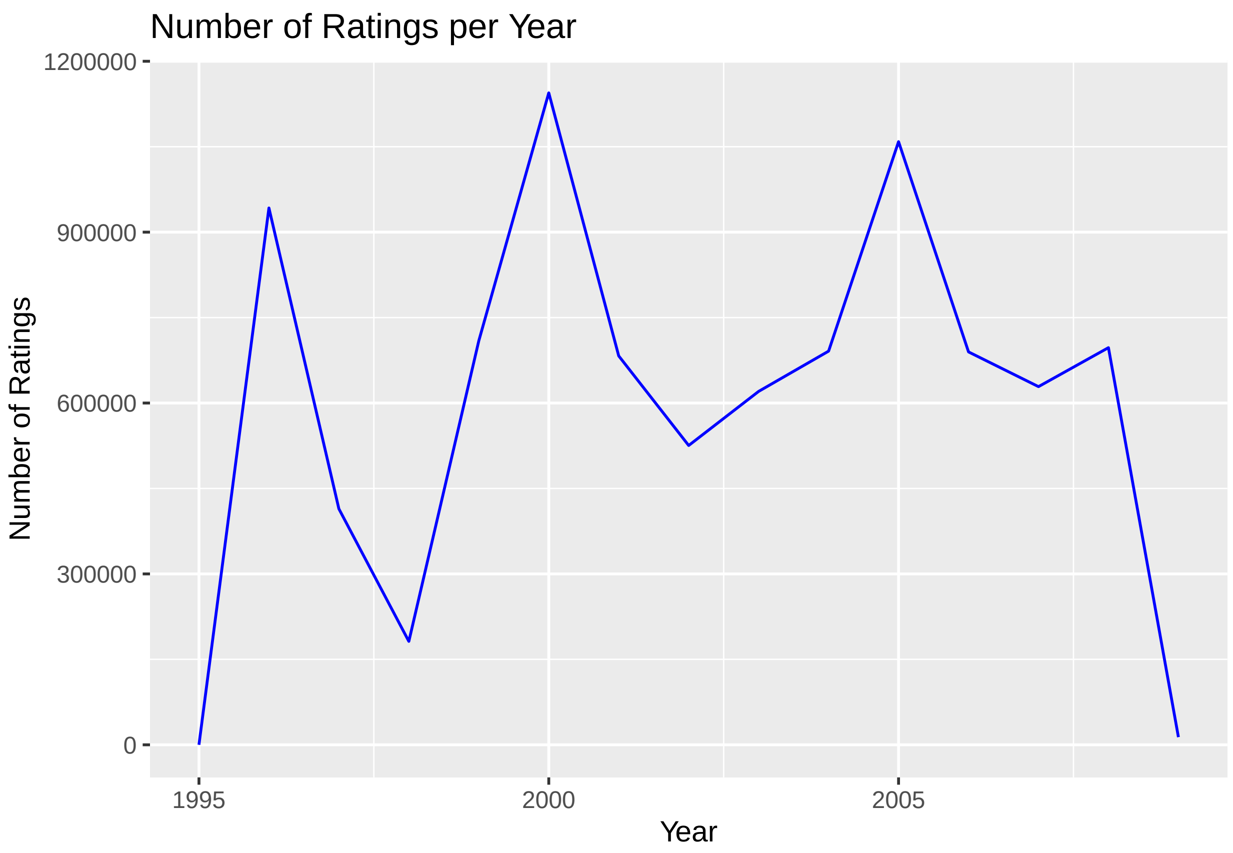
Step 2.3 Data Visualization

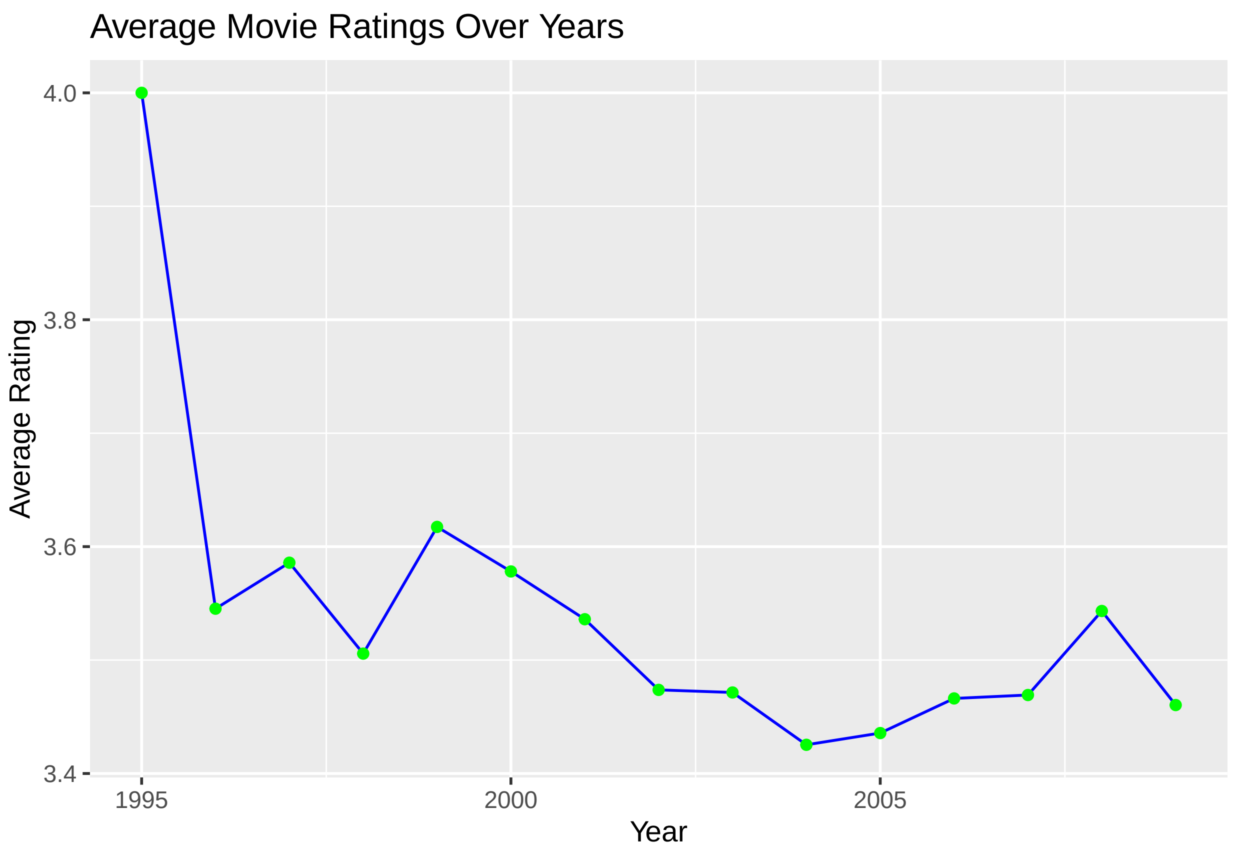
Data visualization is necessary to get more deep insights about the data set. The *distribution of ratings* showed that most users give **whole-number ratings** (4, 3, and 5 are the top three). We observed that **Film-Noir**, **Documentary**, **War** and **IMAX** genre has high rating average but *Distribution of Movies by Genre* shows that these are the **lowest** in count. To observe *Yearly rating trends*, we need to extract the year from timestamp to get visual insight regarding the number of ratings growing over time as more users joined the platform. Most ratings were given in year *1996, 2000 and 2005*, with an average rating of *3.55, 3.58 and 3.44* respectively. The highest average rating is 4 in year 1995 with total count of 2 users, means only 2 unique users gave movie rating of 4.




Distribution of Movies by Genre
Genre
Drama
Comedy
Action
Thriller
Adventure
Romance
Sci−Fi
Crime
Fantasy
Children
Horror
Mystery
War
Animation
Musical
Western
Film−Noir
Documentary
IMAX
(no genres listed)
0
1,000,000
2,000,000
3,000,000
4,000,00
Number of Ratings


Average Rating per Genre
Genre
Film−Noir
Documentary
War
IMAX
Mystery
Drama
Crime
(no genres listed)
Animation
Musical
Western
Romance
Thriller
Fantasy
Adventure
Comedy
Action
Children
Sci−Fi
Horror
0
1
2
3
4
Average Rating


Number of Ratings per Year
1200000
900000
600000
300000
0
Number of Ratings
1995
2000
2005
Year



Step 2.4 Modeling Approach

To predict user ratings, we need to create the modeling approach withe the desired low RMSE value. We'll build multiple models incrementally, starting with simple baseline model, we progressively improved the model by incorporating additional effects and regularization. We'll use RMSE as the evaluation metric. Before building models, define a function to calculate RMSE, which measures the average magnitude of prediction errors.

The simplest model we can build is one that predicts the global average rating (the average rating of all movies) for all user-movie pairs. The formula used is:

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

With $\hat{\mu}$ is the mean and $\varepsilon_{i,u}$ is the independent errors .

```
## Baseline RMSE (Global Average):  1.060331
```

The RMSE on the edx test set comes out to be **1.06**, which is far more from the required target value and it also indicates **poor performance** of the model. From the course (8th Module Machine Learning), we have learned that RMSE can be improved by adding various effects. To the base model we just defined above, we will improve our RMSE by adding movie effects at first, then user effects and lastly genre effects to see how much it improves the model.

We will use the equation:

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

Where: - hat{mu}: The global average rating. - b__i: Movie effect (the deviation of a movie's average rating from the global average). - b__u: User effect (the deviation of a user's average rating from the global average). - b__g: Genre effect (the deviation of a genre's average rating from the global average). - varepsilon__{i,u,g}: Residual error. We will calculate these effects step by step.

```
## RMSE for the movie effect model:  0.9423475
```

```
## RMSE for the Movie + User effect model:  0.8567039
```

```
## RMSE for the movie + user + genre effect model:  0.8563595
```

Till now, we have observed the following through our modeling process: 1. Baseline Model RMSE = *1.060331* 2. Movie Effects Model RMSE = *0.9423475* 3. Movie + User Effects Model RMSE = *0.8567039* 4. Movie + User + Genre Effects Model RMSE = *0.8563595*

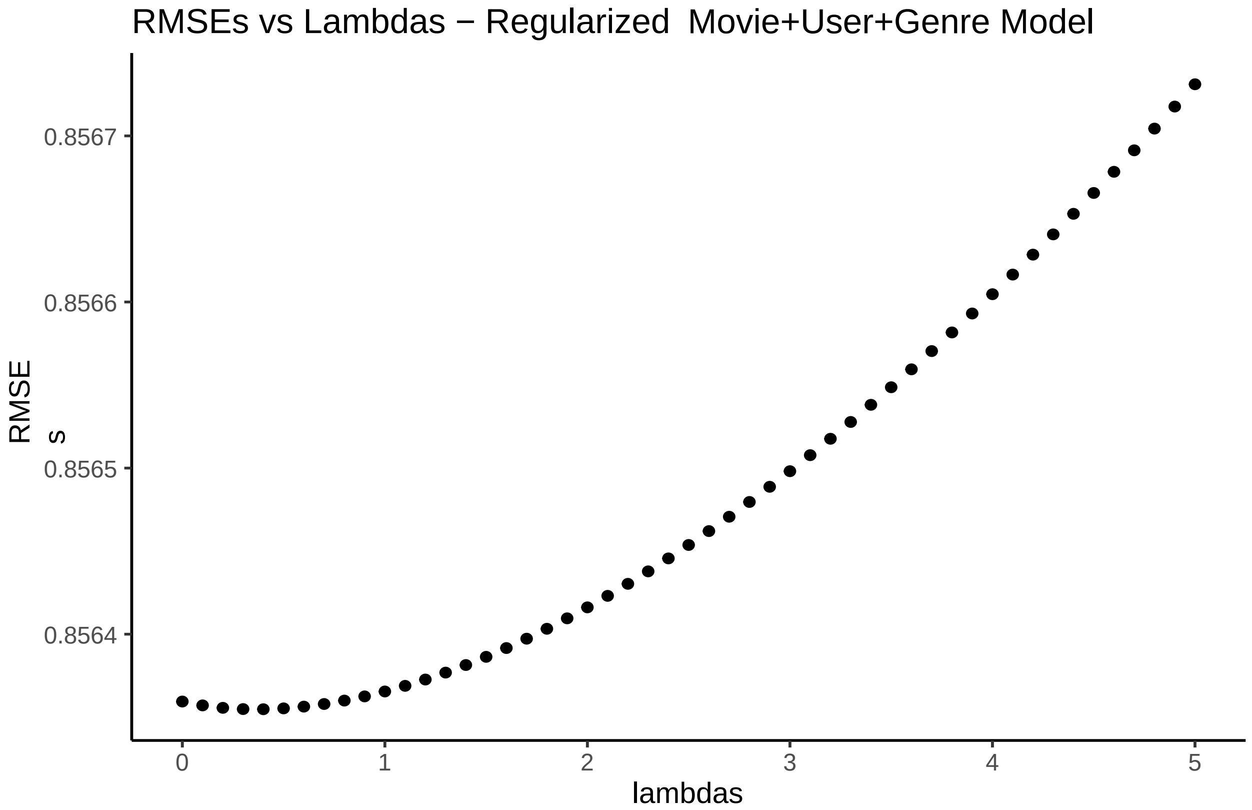
It is evident from the process that, Movie effects and User effects have greatly impacted the performance and reached our desired outcome. The required RMSE value should be *less than 0.8649* and we have achieved *0.8567* by using Movie + User effects Model. However, adding Genre predictor does improves the model slightly to **0.8563**. If we apply regularization technique to this model (Movie + USer + Genre), we may get more improved version.

Step 2.5 Regularization

The regularization method allows us to add a penalty λ (lambda) to penalizes movies with large estimates from a small sample size. When the sample size is very large, the estimate is more stable, but when the sample size is very small, the estimate is shrunken towards 0. The larger the penalty parameter λ, the more the estimate is shrunk. As λ is a tuning parameter. We will uses the following equation:

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

Note: RMSE function will take some time to process.



RMSE for the Regularized movie + user + genre effect model: *0.8563548*

```
## RMSE for the Regularized movie + user + genre effect model:  0.8563548
```

```
## Lambda Value:  0.4
```

3. Results

Here's the summary of the modeling process: 1. Baseline Model: Predict the global average rating for all movies and users. RMSE: **1.060331**

2. Movie Effect Model: Incorporate movie-specific effects by considering deviations of individual movies from the global average. RMSE: **0.9423475**
3. Movie + User Effect Model: Add user-specific effects to the movie effect model, capturing how users deviate from the average. RMSE: **0.8567039**
4. Movie + User + Genre Effect Model: Adding Genre effects to the movie and user effects model. RMSE: **0.8563595**
5. Regularized Movie + User + Genre Effect Model: Apply regularization to penalize extreme estimates of movie, user and genre effects, improving generalization. Optimal Lambda: **0.4** RMSE: **0.8563548**

Regularization was applied to avoid overfitting, especially for movies or users with very few ratings. We used penalized least squares, introducing a penalty term to shrink the effect estimates. The regularized model helped prevent the model from giving extreme predictions for movies or users with low representation in the data set. The optimal lambda for regularization was found to be 0.4, and the model performance improved very slightly as a result.

```
## Best Model Name:  Regularized Movie-User-Genre-Based Model

## Best RMSE value:  0.8563548

## Best Lambda Value:  0.4
```

3.1 Final Evaluation on Holdout Test Set

The final model was evaluated on the holdout test set, which contains 10% of the data that was not used during model development. The final RMSE on the holdout set was **0.86484**, indicating that the model has the best performance and lowest RMSE as required by the course.

```
## Final RMSE value on Final Holdout Test Set:  0.8648472
```

4. Conclusion

After training different models, it's very clear that *movieId* and *userId* contribute more than the *genre* predictor. Without regularization, the model can achieved and overtook the desired performance, but the best technique we found is applying regularization and adding the *genre* predictor, which became the best result for the trained models. Through this approach, we were able to improve the prediction accuracy, achieving an RMSE of **0.8648472** on the final holdout test set.

4.1 Limitations

1. We did not explore advanced techniques such as matrix factorization, which may further improve accuracy.
2. As the data set is very large, we are unable to use kNN or random forest modeling techniques. These techniques require more computational resources, and for such a large data set, the system requires dozens of memory to make computations.
3. The data set seems to be limited to less variables for observations and insights, such as movie ID, user ID, genres combined. If more variables are added such as age and gender, may give more insights about a group and how they rate movies based on which genres.

4.2 Future Work

1. Reduce the data set through sampling technique to apply different models such as kNN and random forest.
2. Matrix factorization, principal component analysis (PCA) and singular value decomposition (SVD) may also be implemented with a sampled smaller data set, to get more insights and to train more models for more accurate results.
3. Adding of more variables such as age and gender, to get more insights about the behavior of rating movies.
4. Addition of separate genres may improve the analysis process.