A Movie Recommendation System Using MovieLens Data

I. Introduction

Today machine learning is widely used in many industries. For instance, companies like Amazon use it to explore consumer behavior and target their products to a specific audience. A movie recommendation system is one of the applications built on top of machine learning algorithms. It is used to predict what rating a given user will assign to a movie, and ultimately it aims to provide users with movie recommendations based on their preferences and viewing history.

This report describes how I built a revised movie recommendation system based on the initial model discussed in the Edx Machine Learning course. The data used to train this system comes from the MovieLens 10M dataset, which was collected by GroupLens Research. It includes 10 million movie ratings and is publicly available for free download. The revised movie recommendation system introduces two new factors, the genres effect and the time effect, in addition to the movie and user effects of the original model. The performance of the new model is then evaluated using the RMSE score. It is compared against the RMSE of the original model with the goal to reduce it to be less than 0.8649.

II. Methods and Analysis

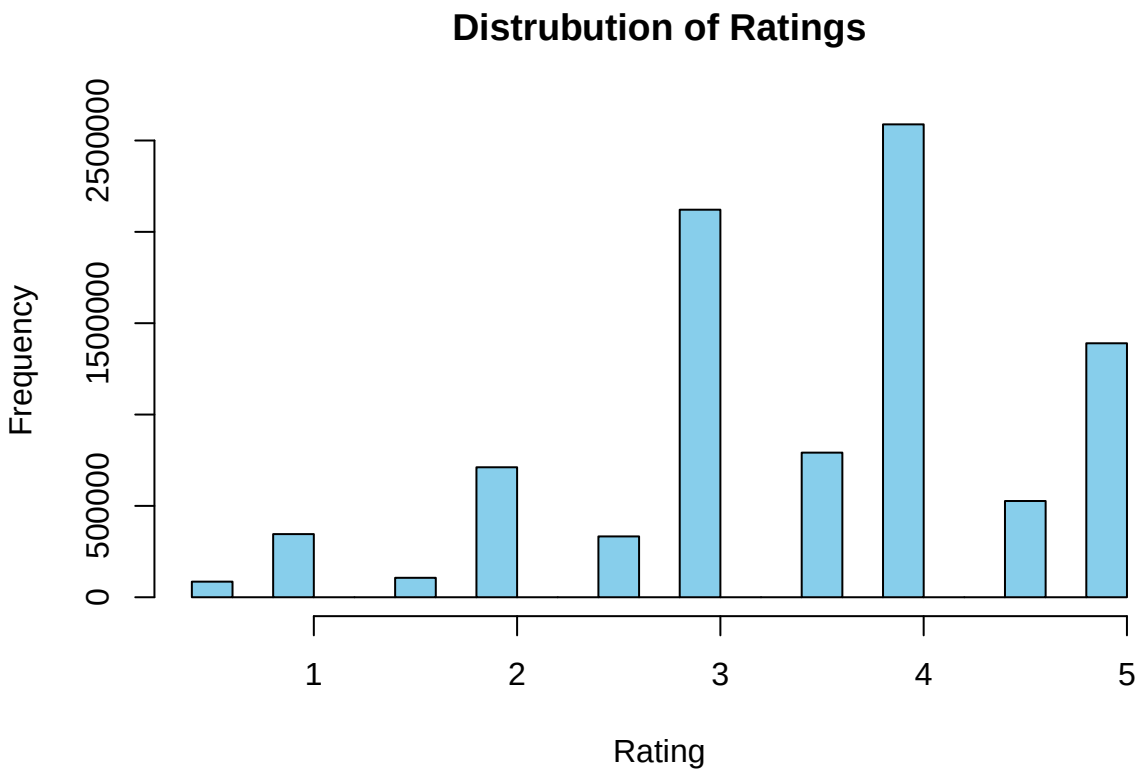
1. Data Processing and Exploration

GroupLens Research has collected various movie rating datasets and make them available on the MovieLens web site. The 10M version of the MovieLens dataset was selected for this project instead of other versions for computation purposes. The MovieLens 10M dataset can be found here.

The 10M data was downloaded from the MovieLens web site and then partitioned into two datasets, the edx dataset that was used to create the new model and the validation dataset that was used as the final hold-out test set to evaluate different algorithms. The validation dataset was supposed to be 10% of the entire 10M dataset. However, in order to make sure users and movies in the validation set were also in the edx set, an additional step was performed. As a result, the final count was 999,999, not 1,000,005.

A preliminary review of the edx dataset indicates it contains roughly 9 million records (N=9,000,055) with almost 70,000 unique users (N=69,878) and over 10,000 movies (N=10,677). It also includes movie ratings ranging from 0.5 to 5.

Here is the distribution of movie ratings in the edx dataset.



And here is a sample of the data itself. We can see it has altogether 6 variables including the 3 variables mentioned above.

```
##   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
```

The timestamp variable in the dataset reports the date and time when a user rated a movie. However, it was not presented in a format easy to interpret, so it was converted using the lubridate package. In addition, a new variable was created based on this variable to reflect the year a movie was rated. With this new variable, I was able to calculate the difference in years between the movie debut year and when it was rated. This value was stored in the yrdiff variable and was used later in the model. Similar data processing steps were executed on the validation set, too.

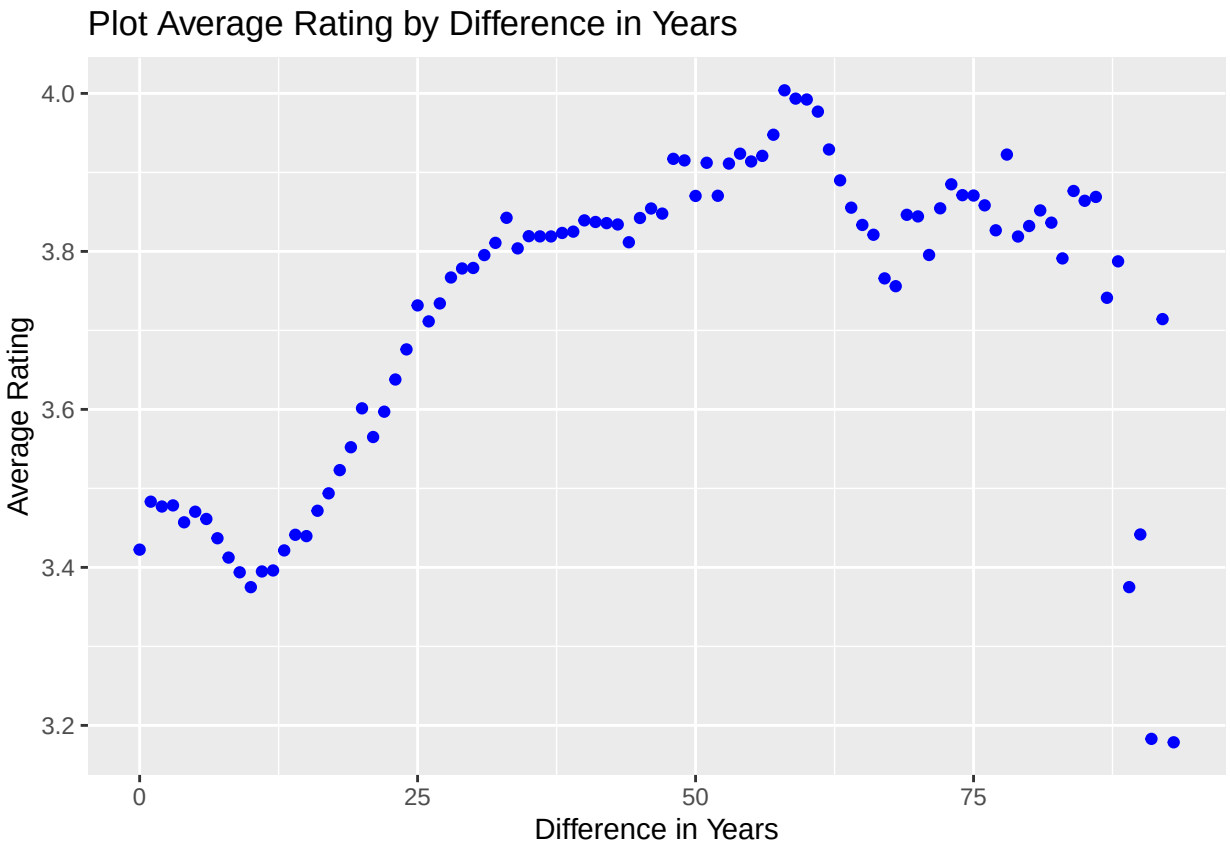
2. Model Development

To build the new model, the edx dataset was further partitioned into two more datasets, the training dataset and the test dataset, with the test data to evaluate the model fit on the training data while tuning model parameters. It's common practice that the test dataset and the validation dataset have approximately the same number of records, therefore the test dataset was created to be 11% of the edx dataset. Again an extra step was performed to make sure users and movies in the test dataset were also in the training dataset. The final count of the test set was 989,991, which was very close to the number in the validation set.

Since the goal of this project was to improve the performance of the original model, it was natural for me to build the original model first and calculate its RMSE as the benchmark. The test dataset was used to find the best value for the regularization parameter given the original model, and then the model was applied to the validation dataset to report the final RMSE from that data.

The next step was to identify additional variables that could be used to enhance the model. When reviewing the training dataset, I noticed some variables other than user and movie that potentially could be candidates for the new model. The first one was genres. As we know, differe movie genres appeal to different audiences. The person who like certain types of movie, for example, comedy, may rank comedy movies higher than other genres. Therefore the genres variable was added to my first model, New Model I. From the Edx course, I learned it was common to use 5 or 10 as the value for the regularization parameter, lamda, so I decided to test this particular parameter ranging from 0 to 10. It turned out that the value of 4.75 gave the minimum RMSE for the model using the movie, user and genres effects. Consequently, this value was used in New Model I, which was then applied to the validation set to calculate the RMSE and compare that to the benchmark RMSE.

The second variable that was added to my model was the difference in years. Movie ratings can change over time. People seem to be critical or harsh on recent movies while they tend to be more lenient with older movies. This pattern can be observed in the following chart.



Given this trend, the time effect, which was measured as the difference in years (i.e. the difference between the movie debut year and the year when it was rated), was introduced to the model, New Model II. Again, a similiar approach was adopted as New Model I, with the test set to identify the best value for the regularization parameter and then the final model was applied to the validation set to compute the RMSE of the model.

III. Results

The original model has a RMSE of **0.8648201** (see below), which serves as the baseline to compare against new models.

```
## RMSE of the original model:  0.8648201
```

The first model (New Model I) introduces a new variable, the genres effect, which is different from the original model that uses the movie and user effects only. This model has a RMSE of **0.8644514** (see below) when run on the validation set. This is lower than the RMSE of the original model and proves that the genres effect does improve the model performance and is in fact a useful predictor for movie ratings.

```
## RMSE of New Model I:  0.8644514
```

The second model (New Model II) adds one more variable to measure the time effect on movie ratings. The final RMSE on the validation set is **0.8640185** (see below). Again this model further reduces the RMSE and improves the model performance quite a bit. Given it has the lowest RMSE among the three, this is the final model that I select to predict movie ratings.

```
## RMSE of New Model II:  0.8640185
```

IV. Conclusion

Movie recommendation systems are quite common in today's world. They are often used to create personalized content for each user and are of particular importance to companies like Netflix. Each system may use different variables and implement different algorithms to achieve the best results. The model discussed in this report is developed based on the original model from the Edx Machine Learning course. It introduces two additional factors, the genres effect and the time effect, in an attempt to improve the model performance. The results indicate both of those two factors are useful predictors and contribute to reducing the RMSE of the final model. Of course these are not the sole factors that may affect movie ratings. User demographics and Movie characteristics are likely to play an important role, too. Even with the existing factors in my model, there are also possibilities for potential improvement. For example, the genres effect can be further examined by splitting the individual genre into separate categories. Future research in those areas can further enhance the movie recommendation system and provide more accurate predictions.