MovieLens Project

Arwa Ashi

09 March, 2022

Contents

Table 1: The total of Unique Movie ID, User ID, and Genres

n_movie	n_user	n_genres
10,677	69,878	797

1 Introduction

There were Netflix recommendation systems competition to reduce the root mean squared error (RMSE). This report is using a relevant dataset to complete a data science course project sharing the same objective of reducing the RMSE. The report is divided into data cleaning and exploration, methodology, result, and conclusion sections.

2 Data Cleaning and Exploration

2.1 Data Cleaning

First cleaning the edx provided data after downloading it from the provided URL. Second, save edx and validation files into csv files to avoid re-doing the fist step. Third, since the edx dataset also includes a timestamp that represents the time and data in which the rating was provided, the units are seconds since January 1, 1970, a new column date with the date was created.

2.2 Data Exploration

The edx provided dataset has 9,000,055 rows and 7 columns, with 69,878 unique users who had rated 10,677 unique movies with 797 unique genres. See Table 1. In the following, exploration for rating, movie id, user id, genres, and timestamp.

After cleaning the data to a stage that can be analyze, data exploration was done. First, exploring the distribution of each movie in the datasets by total and percentage, see Table 2. In edx data, Drama is the most rate it movie with ~ 43% and Romance is the least rated move with ~19% of total ratting. In Validation data, Drama is the most rate it move with ~ 43% and Romance is the least rated move with ~19% of total ratting. However, the total is not sum to the total dataset rows and percentage is not sum to 100 because some movieId share more than one genres.

Table 2: The Total and Percentage of the Movies

movie_type	edx_total_movie	edx_percentage_movie	Valid_total_movie	Valid_percentage_movie
Drama	3,910,127	43.44559	434,071	43.40714
Comedy	3,540,930	39.34343	393,138	39.31384
Thriller	2,325,899	25.84316	258,536	25.85363
Romance	1,712,100	19.02322	189,783	18.97832
Total	11,489,056	127.65540	1,275,528	127.55293

Let's distribute the movie rating overtime see Figure 1 where the initial date for the record is 1995-01-09 and the final date for the record is 2009-01-05. The figure shows that the total rating is different each year.

Lets distribute the unique movies into 797 unique genres and filtering the total movie for which is greater than 300 see Figure 2. Drama movie has the highest total movie type. As a result, the highest movies' genres is Drama and is the most rated by movie id, the second highest is Comedy and is the second most rated by movie id. However, the common genres for user is Comedy then Drama but the users rated Drama the most. see Figuer 3.

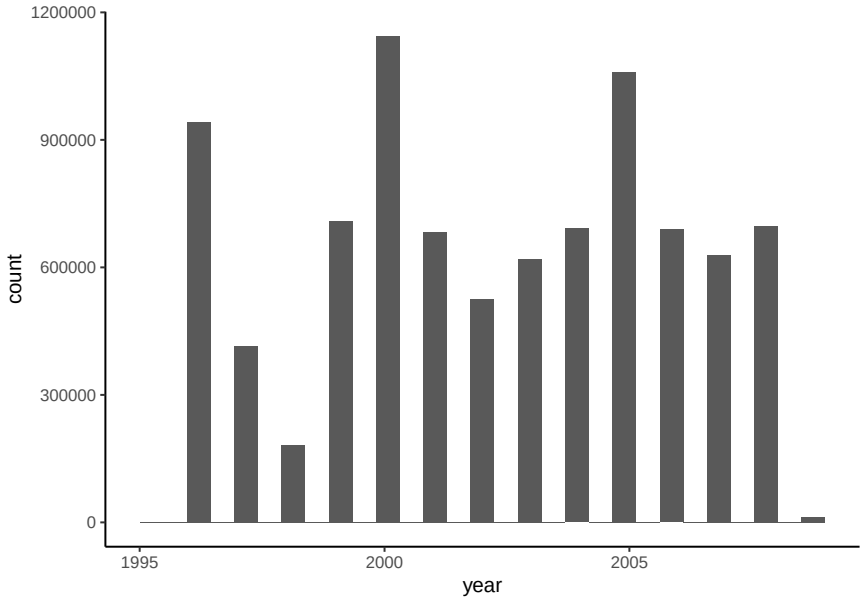

Figure 1: Movie Rating Distrbuted Overtime

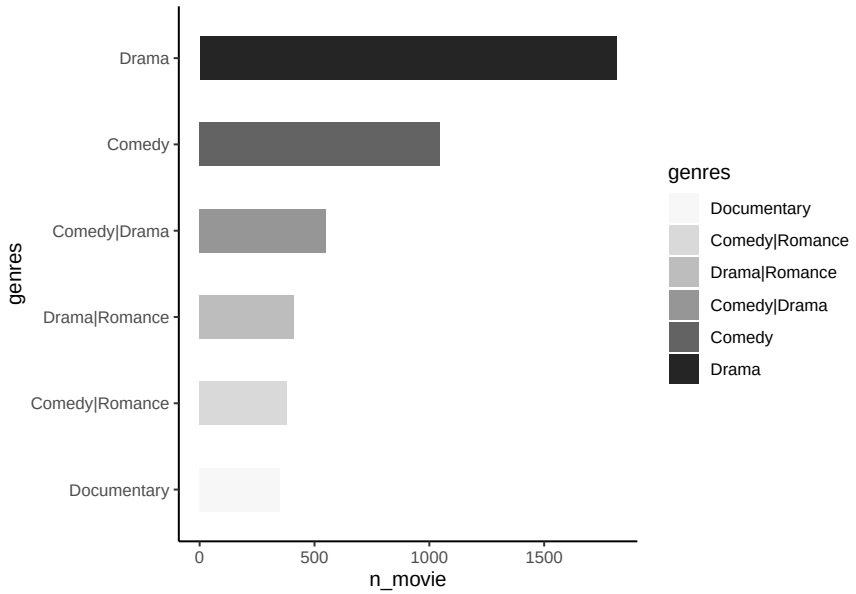

Figure 2: Top Movies' Distrbuted by Unique Genres

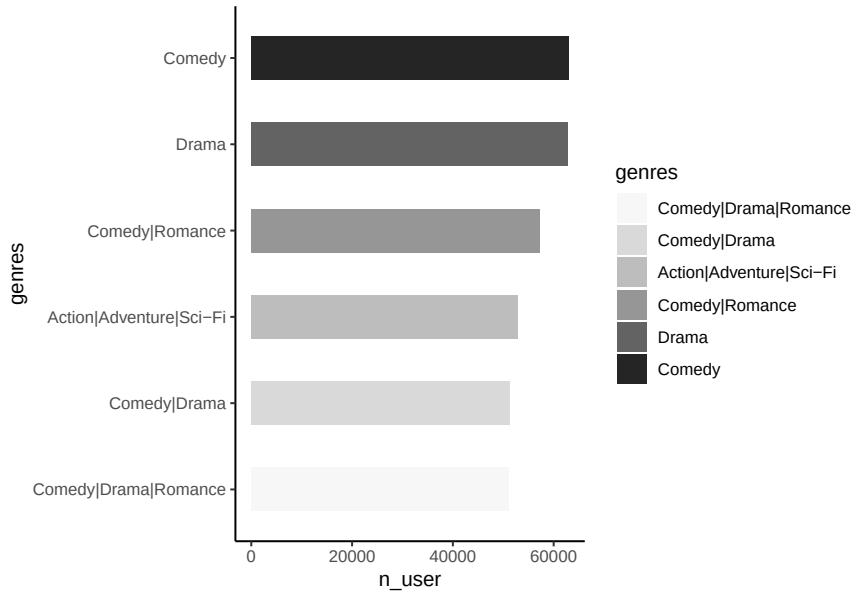

Figure 3: Top users' Distrbuted by Unique Genres

Table 3: The Distribution Over Rating

Var1	Freq
0.5	85,374
1	345,679
1.5	106,426
2	711,422
2.5	333,010
3	2,121,240
3.5	791,624
4	2,588,430
4.5	526,736
5	1,390,114

The average rating is 3.5 over all movies, see Table 3 for the count distribution over rates that started from 0.5 to 5 with Rate 4 having the highest distribution.

As a result, movies has different distribution for each genres and are rated unequally and each user has different total rating movie number. Insights gained, movie, user and genres effects (or bias) need to be considered in the movie recommendation system.

3 Methodology

3.1 Modelling Approaches

3.1.1 Introduction

As a result of the data exploration, movies has different distribution for each genres and are rated unequally and each user has different total rating movie number. Consequently, movie, user and genres effects (or bias) need to be considered in the movie recommendation system. The root mean squared error (RMSE) is considered as a loss function to compare the models to the baseline which is represent in the following

equation,

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

where N is the number of user-movie combinations, $y_{u,i}$ is the rating for movie i by user u, and $\hat{y}_{u,i}$ is predictions.

This report exam several models to improve the RMSE, first model, assuming the same rating for all movies and all users. Second, adding movie effect to the model. Third, adding movie and user effect to the model. Fourth, adding movie, user, genres effect to the model.Fifth, using regularization technique for movie, user and genres effect.

For the validation, the RMSE returned by testing the final algorithm on the validation set (the final hold-out test set).

3.1.2 Splitting The Data Frame

Dividing the dataset into 3 data frame 'Training' has 80% of edx dataset, 'Validation' and 'Testing' has 20% of edx dataset. The Validation data frame provided an unbiased evaluation of a model fit on the training data frame.

3.1.3 Creating and Evaluating the Models

3.1.3.1 First Model We start with a model that assumes the same rating for all movies and all users, with all the differences explained by random variation: If μ represents the true rating for all movies and users and ϵ represents independent errors sampled from the same distribution centered at zero, then:

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

In this case, the least squares estimate of μ, the estimate that minimizes the root mean squared error, is the average rating of all movies across all users.

3.1.3.2 Second Model Improving initial model by adding a term, b_i, that represents the average rating for movie i :

$$Y_{u,i} = \mu + b_i + \epsilon_{u,i}$$

where $b_i$is the average of minus the overall mean for each movie i .

3.1.3.3 Third Model For further improving, adding b_u to the model, the user-specific effect:

$$Y_{u,i} = \mu + b_i + b_u + \epsilon_{u,i}$$

where $b_u$is the average of minus the overall mean for each user i .

3.1.3.4 Fourth Model For further improving, adding b_g to the model, the genres effect:

$$Y_{u,i} = \mu + b_i + b_u + b_g + \epsilon_{u,i}$$

where $b_g$is the average of minus the overall mean for each genres i .

Table 4: RMSE Results

method	RMSE
Just the average	1.0605303
Movie Effect Model	0.9440004
Movie + User Effects Model	0.8440396
Movie + User + Genres Effects Model	0.8434816
Movie + User + Genres Effects Model - validation set	0.8385620
Regularized Movie + User + Genres Effect Model	0.8657234
Regularized Movie + User + Genres Effect Model - validation set	0.8653033

3.1.3.5 Fifth Model For more improving to the results, regularization is considered. Regularization constrains the total variability of the effect sizes by penalizing large estimates that come from small sample sizes. To estimate the b's, we will now minimize this equation, which contains a penalty term:

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

The first term is the mean squared error and the second is a penalty term that gets larger when many b's are large. The values of b that minimize this equation are given by:

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

where n_i is a number of ratings b for movie i.

The larger λ is, the more we shrink. λ is a tuning parameter, so we can use cross-validation to choose it. We should be using full cross-validation on just the training set, without using the test set until the final assessment. We can also use regularization to estimate the movie, user and genres effects. We will now minimize this equation:

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

3.1.3.6 Validation The RMSE returned by testing the final algorithm on the validation set (the final hold-out test set).

4 Result

4.1 Modeling Results

Based on first model, assuming the same rating for all movies and all users. Second, adding movie effect to the model. Third, adding movie and user effect to the model. Fourth, adding movie, user, genres effect to the model.Fifth, using regularization technique for movie, user and genres effect. For the validation, the RMSE returned by testing the final algorithm on the validation set (the final hold-out test set).The RMSE Results see Table 4

4.2 discusses the Model Performance

The RMSE score is improved by adding additional effect to the equation.

5 Conclusion

There were Netflix recommendation systems competition to reduce the root mean squared error (RMSE). This report used a relevant dataset to complete a data science course project sharing the same objective of reducing the RMSE. The report was divided into Data cleaning and exploration, methodology, result sections. The RMSE score is improved by adding additional effect to the equation. The future work is to study the timestamp effect on the RMSE score.