Capstone - Movielens Project

Ammar Thuraya

5/26/2022

Introduction and Executive Summary

Movie-lens is the first project that we cover part of the “HarvardX: Data Science - Capstone Project”: Movie lens.”

We will build a model using R language to predict movie ratings via the provided MovieLens dataset. Through out the project, we will do the following steps:

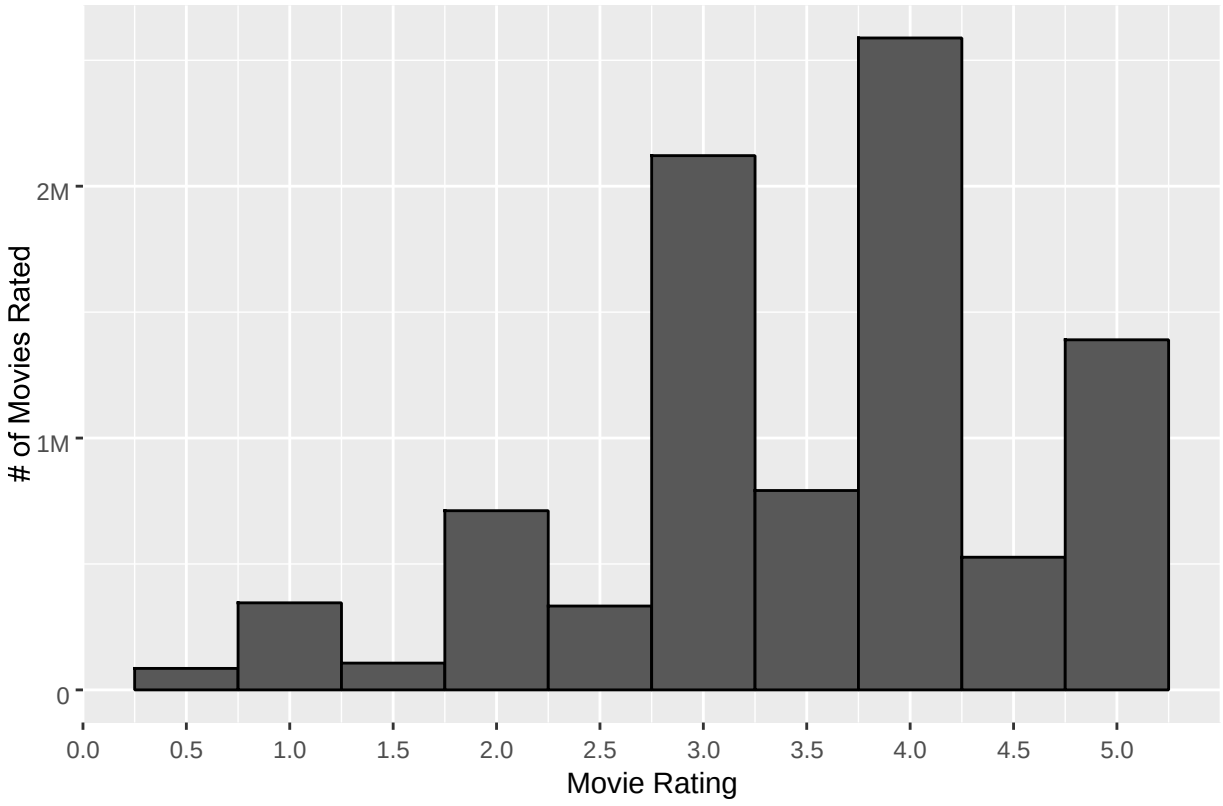
1. Create Edx dataset from the source MovieLens database
2. Explore the dataset and use few techniques do data to do data cleaning, and visualization
3. Provide insights that were gained, and the modeling approachthat was used to achieve the least Root Mean Square Error (RMSE) which will help in providing optimum movie predictions for users.
4. Results section to present the modeling results and discussesthe model performance
5. Conclusion section to provide a brief of the report summary andresults

More clarification on the modeling techniques

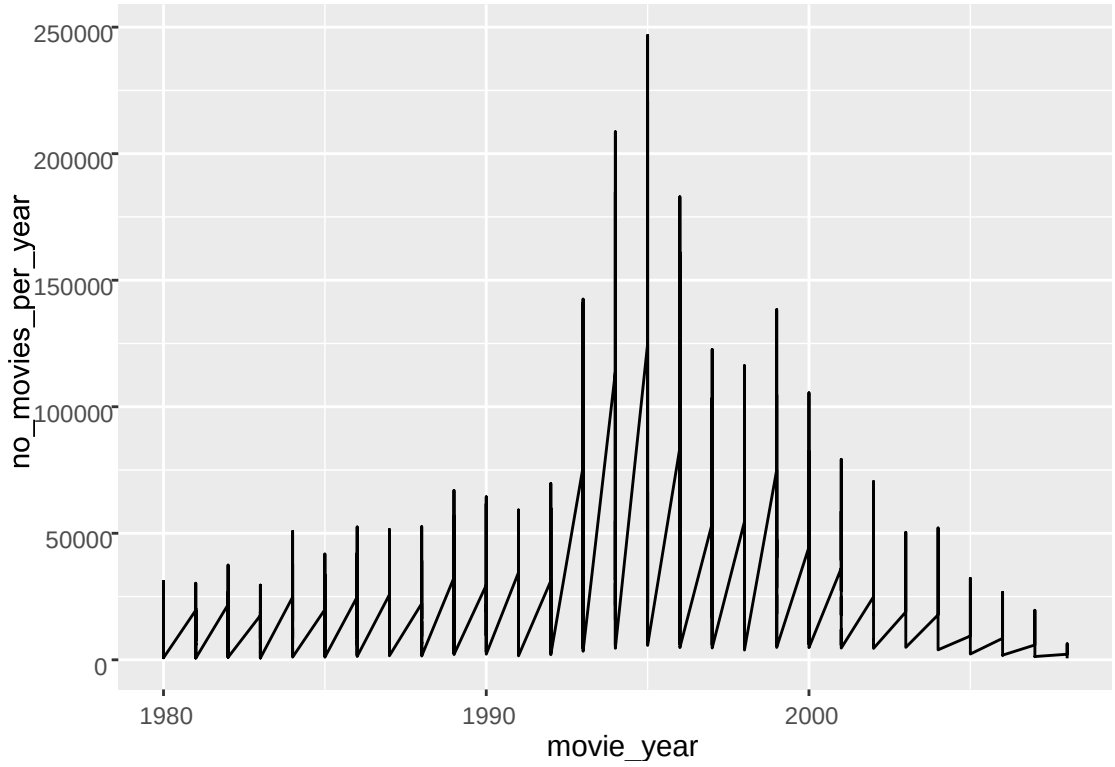
We will use regression analysis a technique to better understand the relationship between one or more predictor variables that could have an effect on the results. To assess the regression model results and how it best fit our dataset, we will calculate the ‘root mean square error’ or ‘RMSE’, for each selected model to measure the average distance between the predicted values from the actual values in the dataset. The lower the RMSE, the better the model can “fit” our dataset.

The dataset

Following chart is for the Distribution of Movie Ratings in the dataset:



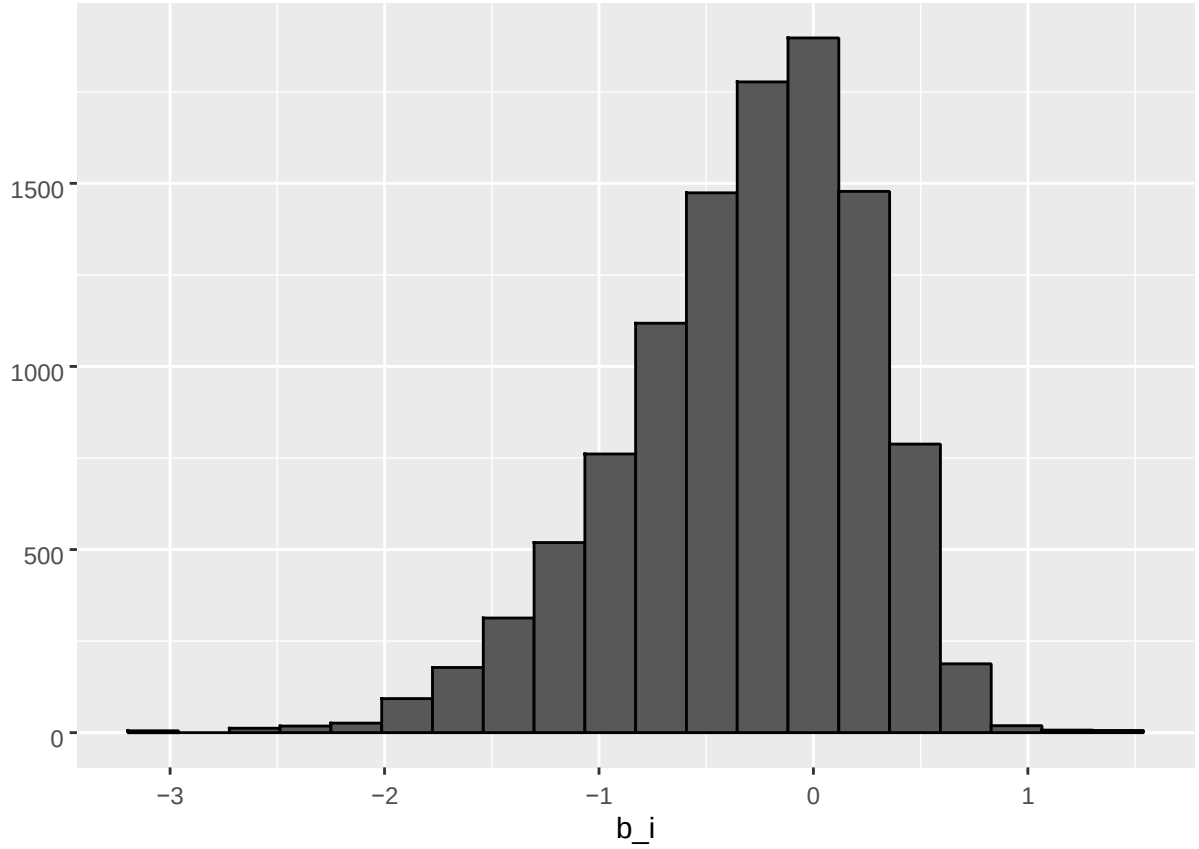
Following is a chart showing the number of movies per year in the dataset:



The Modeling Section

Training Model 1: with Movie effect

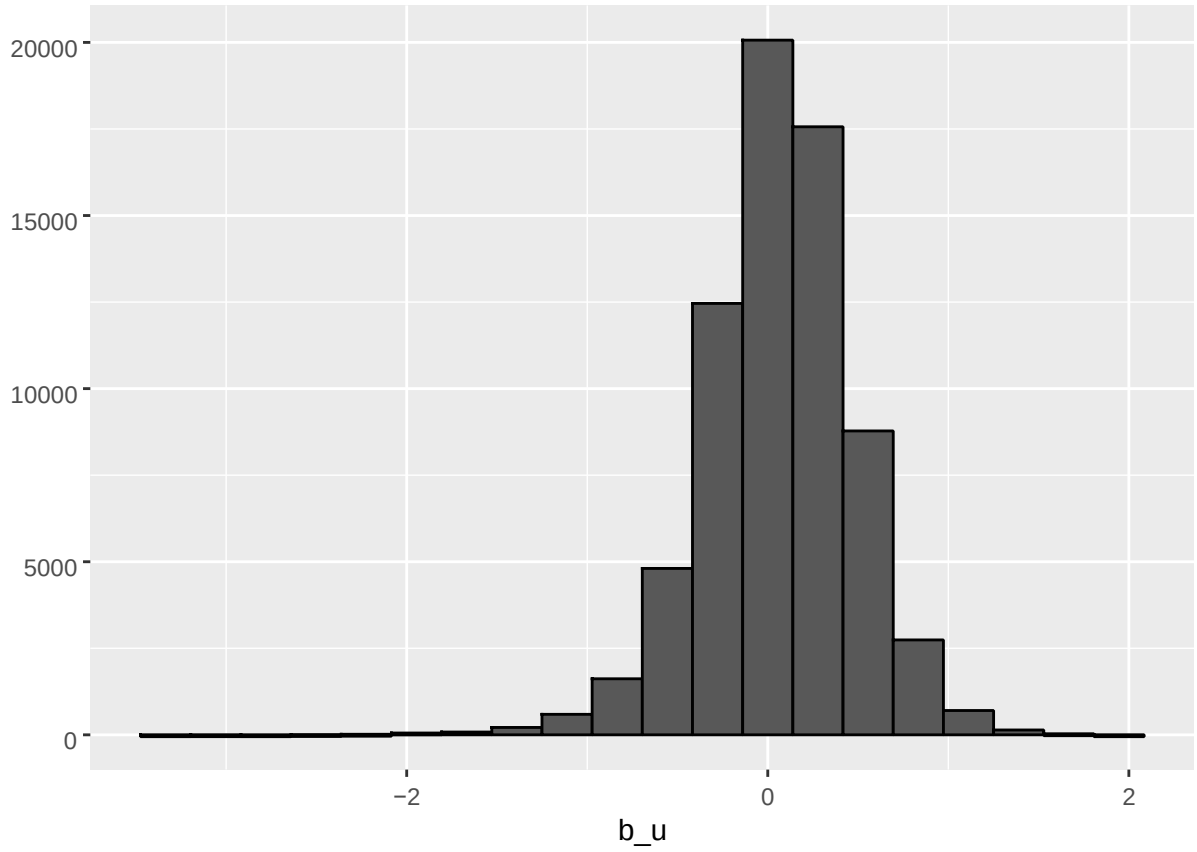
For this model, we use the linear regression model by using the average of all movie ratings. In the model, we will use the average of movie ratings as an effect, and add the term b_i to represent the average ranking for movie i.



Using Model 1, we were able to get an RMSE value of 0.9439.

Model 2: Measure the User effect

For Model 2, we use the average user ratings as the predictor, with bias effect b_u, and get the difference between user average ratings and the average of all users movie-ratings.



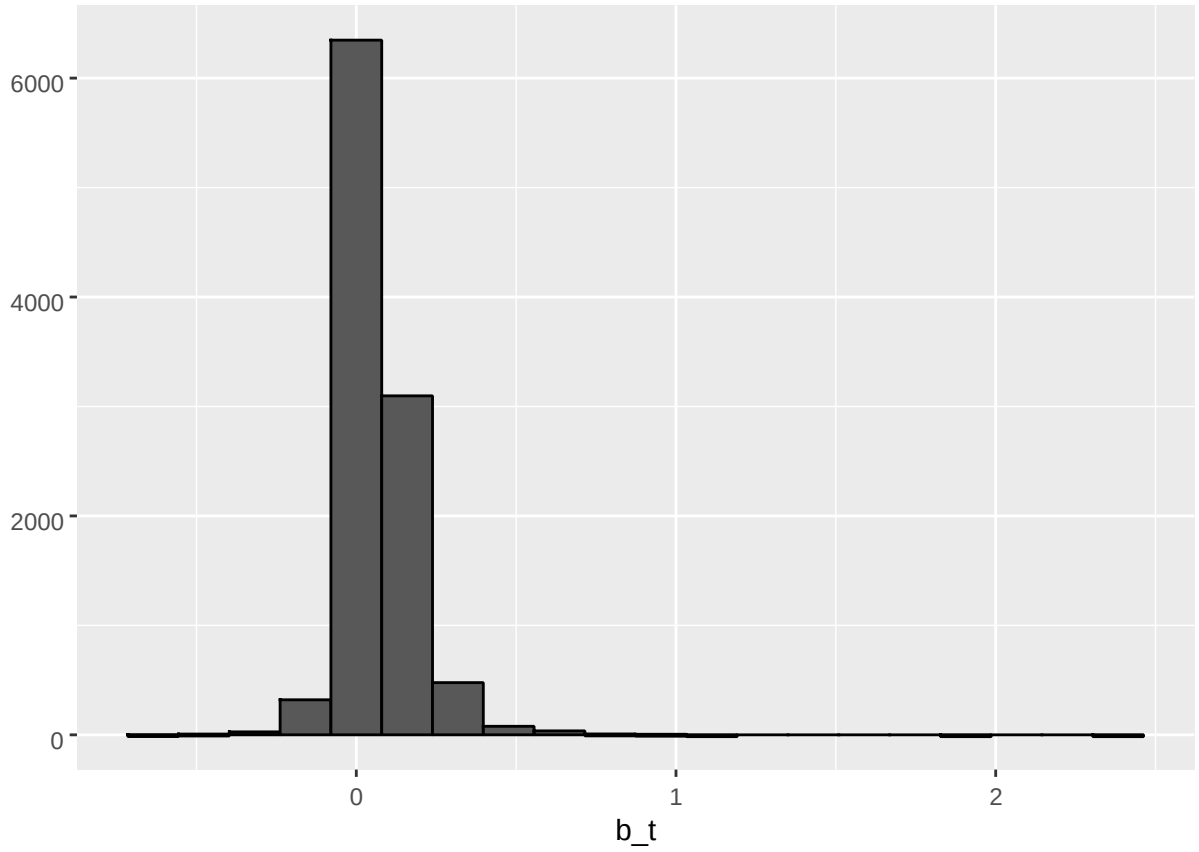
Using Model 2, we were able to get an RMSE value of 0.9948.

Model 3: the Movie and User effects

In Model 3, we use both the movies and users' effects, with two biases b_i and b_u to test this model. Using both effects help show some improvement and were able to get an RMSE value of 0.8653, but was not sufficient enough to get the desired results and outcomes.

Model 4: modeling for the three effects movies, users, and title

In Model 4, we used three effects: movies, users, and title all together. We used the biases: b_i, b_u from Model 1 and Model 2 above, in addition to the new bias b_t for titles. The outcome of this model was better than the previous ones, but it still did not help in reaching the desired outcomes. Thus, the need to regularize and tune the models further



Using Model 4, we were able to get an RMSE value of 0.8641.

Using Regularization and Cross Validation for further improvements

'Regularization' is a technique that would allow us to penalize large estimates that are formed using small sample sizes. This method will help to constrain the total variability of the effect sizes like movie, user and title ratings. We will use cross-validation on the training set, and use the testing set in the final assessment. We will not be using the testing set for tuning.

lambda will be our tuning parameter, and we used cross validation to select it.

We used this method to regularize and fine tune the best model from the above models "Model 4", and we were able to get an optimum RMSE of 0.86377 that meets the requirements of the projects.

The Results and Findings

In this project, analysis and exploration was done on a large 10MB set called 'edx'. Then, data modeling was done on the edx set using a training set (edx) and a validation or 'testing set' that was was a subset (around 10%) of the training set. Four different models were tested to find the best fit model with the least RMSE. RMSE (or the 'Residual Mean Square Error') measures the typical error that can be made when predicting the movie rating. The five models that were selected in this project measured the RMSE values using different combinations of bias effects: movie, user, and title ratings. To improve the accuracy of the predicted results, we further enhanced the best model RMSE further by adding the lambda tuning factor to regularize the results.

Following is a summary table of the findings:

Model 1 RMSE	Movie effect	0.9439
Model 2 RMSE	User effect	0.9948
Model 3 RMSE	Movie+User effects	0.8653
Model 4 RMSE	Movie+User+Title effects	0.8641
Model 5 RMSE	Regularized Movie+User+Title effects	0.86377

Conclusion

In this project, we have identified the optimum or best fit model that provides an RMSE lower than 0.86377 as requested in the assignment. Further tests and tuning of the models can be done by measuring and introducing other effects like gender and time which were not covered in this project.