Overview of the Project:

For the MovieLens Project, I attempted to create a movie recommendation system that draws from predictions made on the MovieLens dataset. This dataset includes ratings for movies and users and other information about movies, which allowed me to predict the rating that a movie would garner. For this, part of the dataset was used for the machine learning task, so that the rating prediction system can be created, while the other part of the dataset was used to test the predictions of the code and compare them to the actual predictions. This was done using the RMSE, to see how far off the predictions were to the actual dataset. To do the machine learning task, I found that Penalized Least Squares was the best method to make predictions and to compare the values to the actual dataset.

Methods/Analysis:

At first, I looked through the dataset to see exactly which values the dataset would have and to think of ways that I can perform the machine learning task for the best results. I first looked at potential machine learning solutions I could have used to generate and accurate prediction model. After comparing the Distributed Random Forest method and other methods, I found that the Penalized Least Squares method would be the best choice for the analysis. I then proceeded to see which columns and data points would be best to factor into the prediction. I saw that the userID and the movieID would be the best decisions to factor into the Penalized Least Squares algorithm and proceeded to make a prediction for the reviews. Since I was taking the users into account for the prediction, I did acknowledge that there would be a discrepancy in the movie rating as some users may not have rated enough movies to properly affect the data. Some users that have not made too many predictions may cause variance as their ratings could fall in the extremes and not properly represent the true average rating for the movie. To factor for this, I took into account a penalty value that standardized the predictions to decrease the difference in the errors.

Results:

I ran the penalized least squares task on a variety of different penalty values to obtain an optimized Root Mean Square Value and improve the prediction of the ratings. My penalty value was 0.1 and with this, I made the predictions to compare with the other movie reviews. With taking into account the users and the movies, I was able to get the Penalized Least Squares prediction to make predictions with a Root Mean Square value of 0.8567.

Conclusion:

All in all, the goal of the project was met and the RMSE value was able to be comparatively low, making for a good prediction when comparing it to the actual values of the data. The output predictions were fairly accurate and in sync with the actual ratings that the dataset had showed, as seen in the RMSE value around 0.85.