DS Capstone Report: Netflix Project

John Hanratty

July 21,2020

Introduction

This Capstone project focused on designing, organizing, building, testing and documenting a rudimentary movie rating prediction engine based on a particular user's historical preferences and the preferences of the overall population. The EDX data set that was provided for the project contains about 100,000 movie reviews over the period of 1995-2015. Each review identifies the reviewer (user), the movie, a rating assigned by the reviewer (scale 0-5), and genre classification(s) for the movie. Given the short time allotted for the capstone, additional data was neither utilized nor needed to attain the target performance metrics. The project was my first attempt at handling end-to-end development, data management and testing at a modest scale. Learning new processes and tools were as challenging as the data science.

The EDX data set required some study and massaging to enable analysis. In general, the data set was pretty clean. Some minor data grooming was required including extracting / formatting dates and massaging genre information for more convenient processing. Subsequent sections discuss the data analysis and various modeling options that were identified and considered in the project.

Although the RMSE goals set for the capstone project were attained, there is enormous room for improvement of the model. The model predicted ratings of 3 and 4 pretty well but over estimated the other rating levels, especially 5. The half-point ratings (i.e. 2.5, 3.5, 4.5) were underestimated largely because training data before 2002 did not record these ratings. When the model was tested with only data from after 2002 when collection half-point ratings began, the result was a much better prediction of half-points and a significant RMSE decrease from .8627 to 0.8167. Graphs are provided in the Results section below. An investment in converting pre-2002 data might be questionalbe since fashion and entertainment preferences have changed significantly in the last 20 years. :)

The project provided a great hands-on experience. I look forward to more interesting data science projects.

Methods/Analysis

Data Cleaning

The EDX data set provided for this capstone project contains about 100,000 movie reviews submitted from 1995 to 2009. The data fields included movieId, userId, movie title, genres, and rating. To prepare for analysis and modeling, the data set was modified to address some field formatting issues. Other translation issues were noted and deferred to exploration, analysis or model design phases. These additional translations usually involved choosing an approach from several that might alter the performance of the final model (e.g. normalizing or regularizing ratings). The list below summarizes data cleaning issues with -a- indicating modifications to the base sets, and -b- indicating issues deferred for later phases

-a- The timestamp was converted into a human-readable and lubridate compatible format.

-a- The movie release date was extracted from the title and put into a separate numeric field.

-b- In small number of cases, the review data occurred earlier than the movie release. These were noted and deferred to later phases.

-b- The genres had some ambiguities (e.g. Sci-Fi vs. SciFi) that affected string matching. This was handled as part the genre module of the model

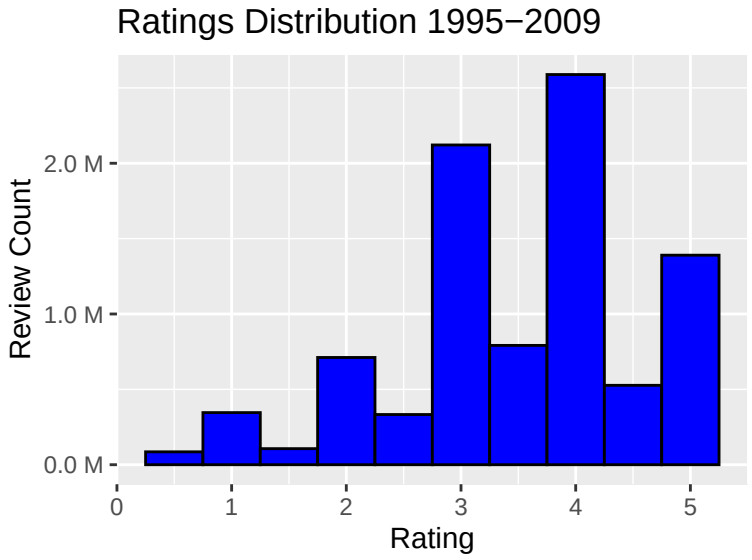
Data Exploration (visualization)

The EDX data set has some interesting characteristics and opportunities for analysis. A few of these are explored in this section.

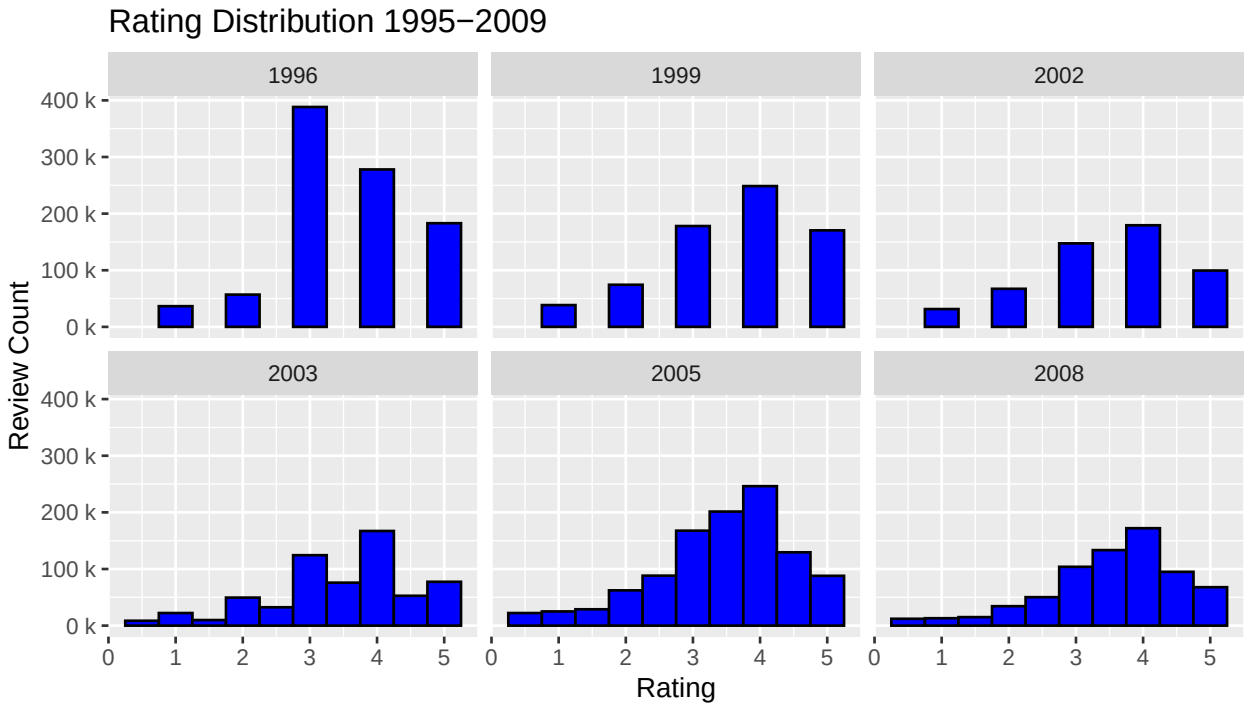
Rating Distribution for Data Set

The graph below shows rating distribution for all reviews from 1995 – 2009 and exhibits a some notable characteristics:

- the distribution seams skewed to the right with a 4 rating being most popular.
- the half-point ratings (e.g. 2.5, 3.5, 4.5) seem less popular.



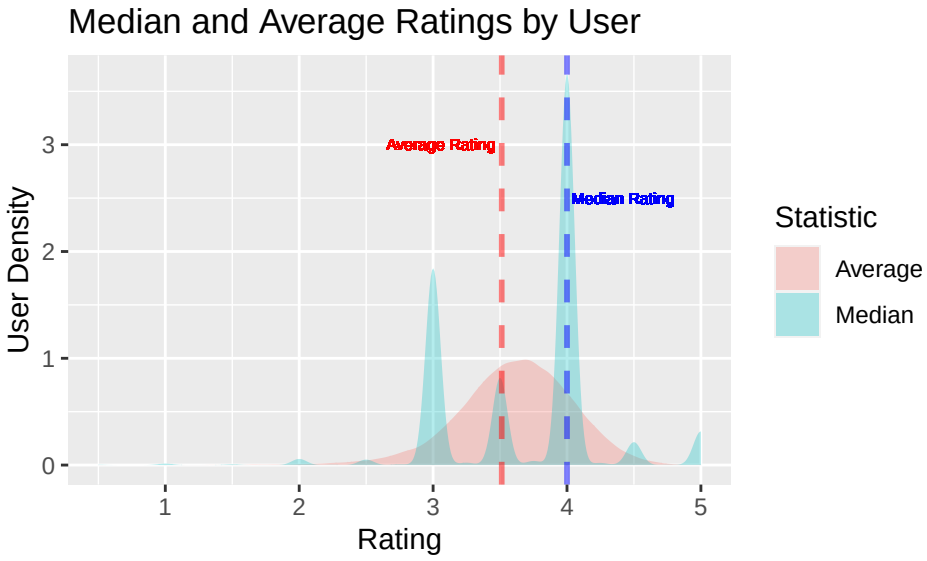
The graphs below break out the ratings distribution by selected years. They show that there are no half-point ratings before 2003. After 2003, half-point ratings exist and seem to have the expected distribution. 2003 was a transition year between the two rating systems.



The above graphs show that 4-points is the most popular rating with a distribution skewed toward higher ratings. Further research is required to fit a distribution curve. In addition, the years before 2003 should be smoothed to fill in the half-point gaps (or replaced by more recent data with half-point ratings). This would allow a better comparison for all years.

Ratings Disribution by User

Each user has their own unique scale for rating movies. This makes it difficult to compare movies and make accurate recommendations. The graph below shows the distribution of the average and median ratings rewarded by users. The previous section shows that 80% of ratings fall within a 2-point rating window (3-4.5). The graph below show that user rating averages vary by as much as 1.5 points on a 5-point scale.



It's impossible to get all users to adopt the same rating scale retrospectively and in the future. To produce an accurate prediction model, a single rating scale is needed. The prediction model requires a translation to a standard scale for analysis. The model can then translate the prediction to account for the individual user.

Genre breakdown

One promising prediction model input is user preferences by movie genre. This information might provide relevant information for matching users with movies. In the EDX data set, movies are assigned one or more genres. The table below shows the frequency of movies genre classifications.

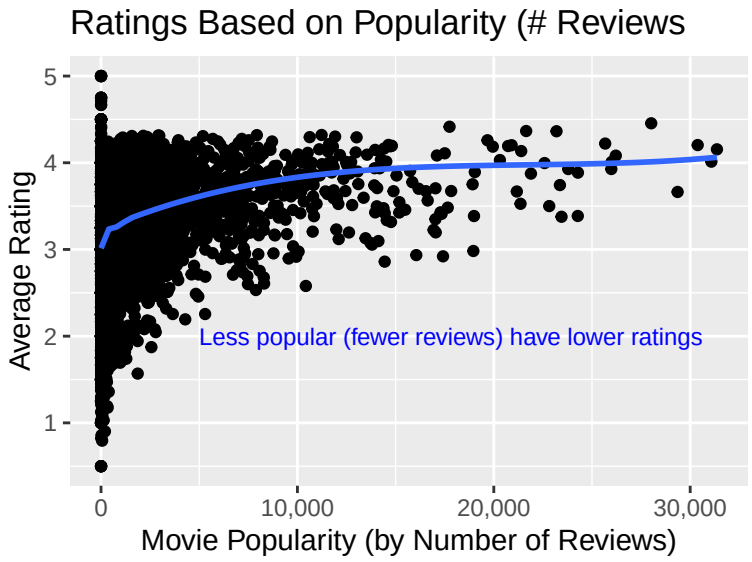
Table 1: Number of Movies with Each Genre Designation

Genre	Movies
Drama	3,910,127
Comedy	3,540,930
Action	2,560,545
Thriller	2,325,899
Adventrue	1,908,892
Romance	1,712,100
Sci_Fi	1,341,183
Crime	1,327,715
Fantasy	925,637
Children	737,994
Horror	691,485
Mistery	568,332
War	511,147
Animation	467,168
Musical	433,080
Western	189,394
Documentary	93,066

A hypothesis is that users will tend to reward similar ratings to movies of the same genre. Further research is required to understand the relationship.

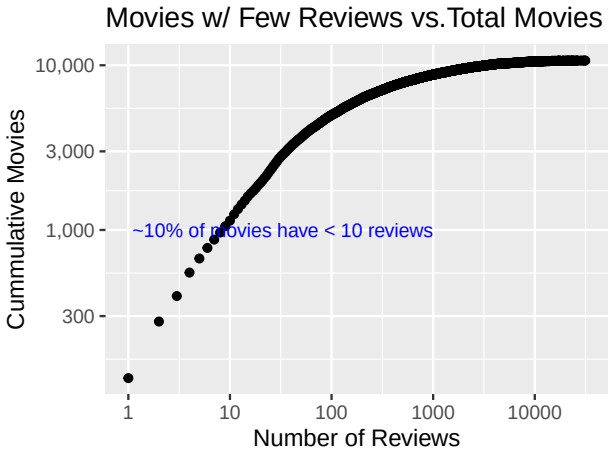
Rating Based Movie Popularity (# Reviews)

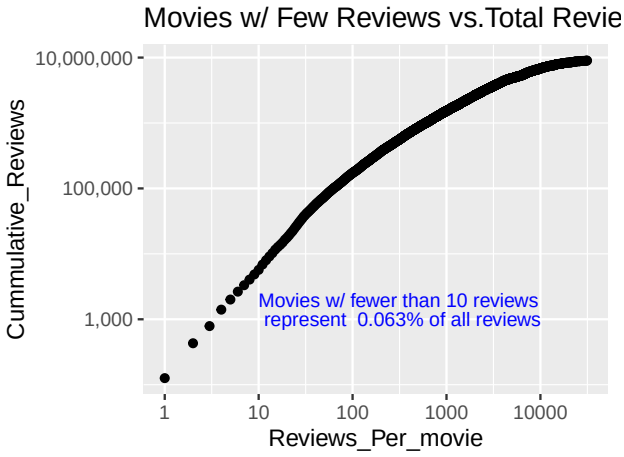
The table below shows that movies with more reviews tend to have higher overall ratings. A hypothesis is that more popular movies are reviewed more often. It seems plausible that this observation could improve predictions.



Reviews per Movie

The number of reviews associated with a movie may indicate popularity but also indicate the quality of the rating average. A large number of reviews tends to cancel the effects of 'outlier' reviews. Movies with only a few reviews have less reliable ratings since one outlier review can substantially affect the average rating. The table below shows that over a quarter of movies have fewer than 10 reviews. This indicates that "regularization" should be considered for the prediction

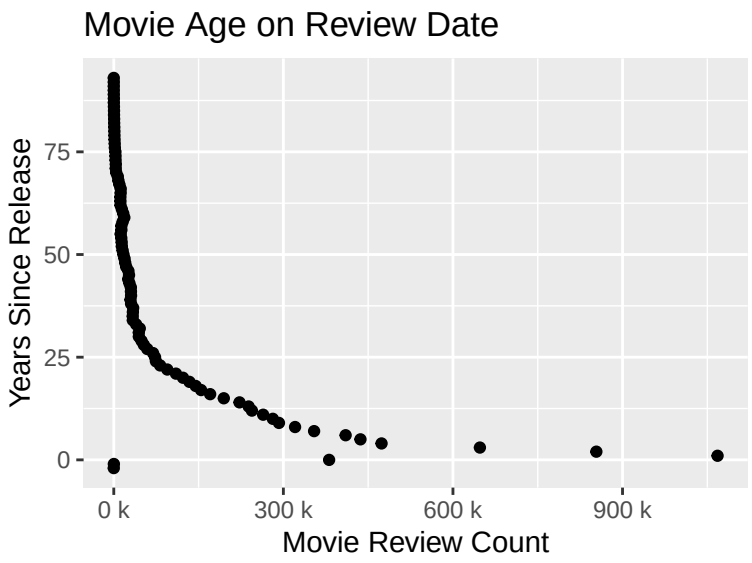




Movie Age

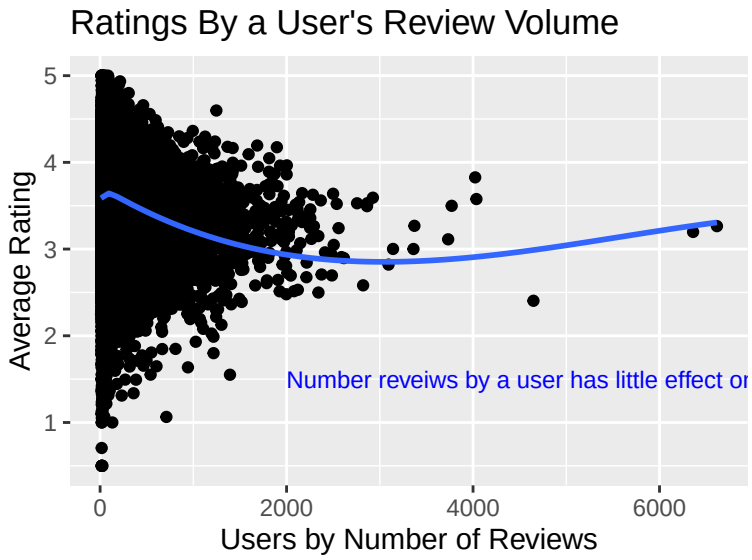
Many users have a preference for older movies, other don't like them at all. In this case, we defined age as the difference between the review data and release date. As seen in the graph below, movie age might

provide a useful indicator for prediction.



Number of Reviews

Some users review a lot of movies. The question is whether this is better or worse predictions of a user's preference.



Data Set Insights

The previous section identifies many opportunities for developing an accurate movie rating prediction model. Unfortunately, this capstone project does not provide enough time for a thorough investigation of all these possibilities. For this reason, priorities were set as follows. Further, analysis, development, and testing will tell whether these were the right choices.

1. Ratings reconciliation – A “Tower of Babble” situation exists where all users speak a different “ratings language” that makes prediction very difficult. The prediction model must collect ratings data using one user’s ratings scale, use the data to create predictions, and the match it to another user in a different ratings scale. Addressing this problem was the first priority for this project.
2. Genre Predictions – Genre information in the data set provides the relevant personalized data for prediction. Building an algorithm that predicts a user’s reaction based on historic ratings for movies of the same genre seemed promising.
3. Regularization – The regularization technique used in class could help account for inaccurate ratings on movies with few reviews. This technique might be more important for a movie recommendation than for a rating prediction. In addition, regularization might alter the ratings for a large number of movies but only apply to a small percentage of review ratings. The regularization effects were investigated in this project.
4. Distribution smoothing – The model must predict “half-point” ratings (e.g. 2.5, 3.5, 4.5) but ther are missing in training data prior to 2003. This will affect the model performance. The process used in #1 above might help alleviate the problem.

The items below have promise to improve predictions in future projects.

• Popularity based on number of reviews

• Movie age/era based on release date

• “Super-reviewer” analysis to see whether users with many reviews help prediction.

Modeling Approach

The EDS data set was divided into train_set (90%) and test_set (10%) collection of reviews. The model architecture followed the model described in the textbook equation.

$$Y_{ugp} = \mu + \beta_u + \beta_g + \beta_p + \epsilon_{\cdot}..$$

A modular architecture wase implemented. Modules were developed to perform particular β functions and plugged together via %>% pipes to pass data between phases. Each module uses the residual training rating results from the previous module (the input – output rating). The idea was to create a pipeline of modules can be plugged in, replaced or re-ordered. Modules were developed and added to the pipeline to hopefully provide incremental improvements. This required a standardized input/output for all modules and a simple multiplexing algorithm for sending all data over one pipe. The training and test dataset are summited at the same time for analysis. Most modules support this paradigm, but some scripting was used near the deadline.

Module1 – Baseline Use the Average

The first module provided a starting point by predicting every review with the average rating for the movie from the training data set (training_set). If the test_set contained movies not in the train_set, the average rating for all reviews in the training_set was used. The function PR1_MovAve_v1 is provide in the submission package performs.

Module1a – Movie Rating Adjusted for User Bias

The 1a module improved on Module1 with a calculation that adds user rating bias (b_u) to a movie's average rating. b_u is the bias (+/-) calculated by subtracting the user's average rating (u_ave) from the average of all users. b_u is positive if the user gives higher ratings than average or negative if less that the average.

The predicted rating is the movie's average rating (m_ave) + b_u to adjust for user bias.

$$Y_u = \mu + \beta_u$$

1. The train_set was used to calculate the following values for USERS: u_ave = average of the ratings in all movie reviews by the user (1/user) USER_ave = average of u_ave for all users (1 per train_set) b_u = u_ave – USER_ave
2. Calculate the following value for each movie (MOVIES): m_ave = average rating for a movie using adjusted ratings by all users
3. A left_join of USERS and MOVIES to the test_set is performed and used to calculate Y (prediction): y_hat = m_ave + b_u
4. A left_join of USERS and MOVIES to the train_set is used to calculate the residual value used by the next stage of analysis residual = rating – Y

Module2 – Genre Preference

The second module uses genres to improve rating predictions. This entails 'spreading' the genres for each review into individual columns for ease of calculation.

Training

The training phase of the analysis creates a USER_GENRE.weights table, which has an entry for each user with a calculated weight (average rating) for each genre. Each train_set review is spread to create a column (UG) for each genre weight. If the review pertains to a genre, the movie rating is added to the appropriate UG column. The data.frame is the grouped by user. The weight is the average rating the user gave to movies for each genre. So, if the user reviewed three Drama movies (ratings 2, 3, 4), the weight would be 3.

Prediction

The test_set is spread to provide genre columns for each review. A left_join adds USER_GENRE.weights to the test_set data set. The prediction is calculated by the average USER_GENRE.weight (UG) for all genre applicable to the movie. So, if the movie is marked with genres G1, G7, and G9. The predicted rating is the average of the UG1, UG7, and UG9 weights.

Y = (UG1 + UG7 + UG9) / 3

The residual that feeds the next module in the analysis is calculated from the train_set.

Results

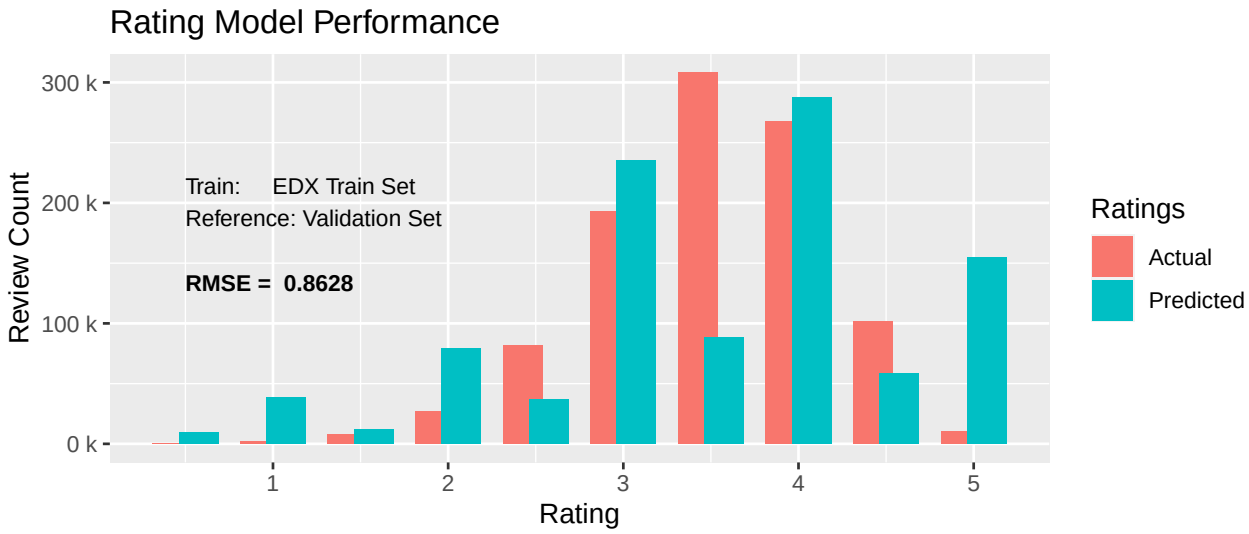
The movie rating prediction model test results are shown on the next page for the both the edx data set and the validation set. While the both meet the RMSE requirement for the capstone project, there is much room for improvement.

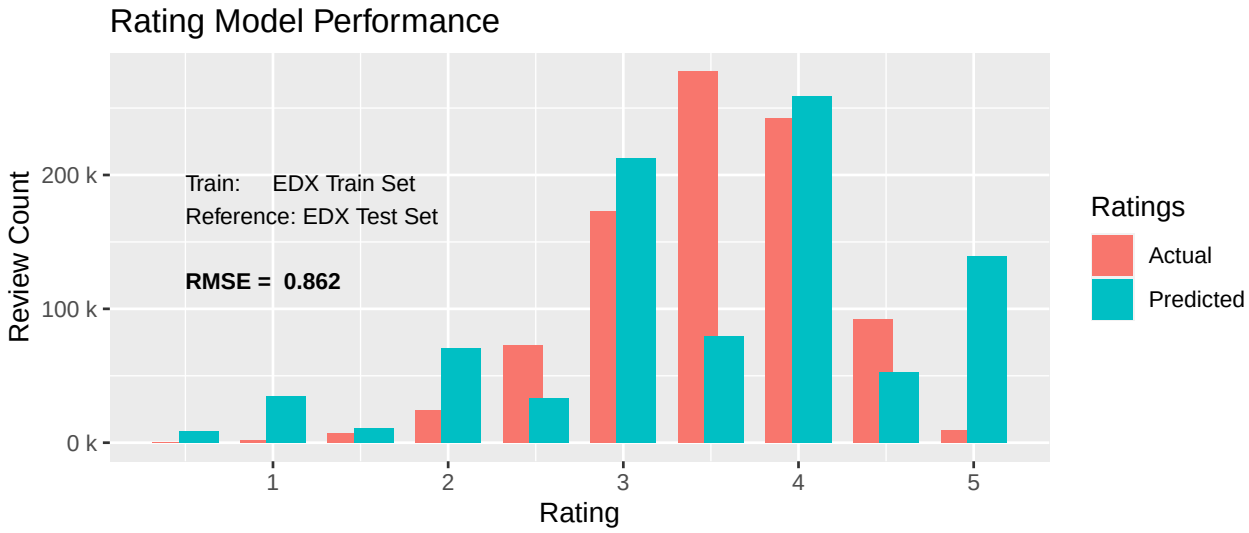
```
## Joining, by = "userId"
## Joining, by = "userId"
## Joining, by = "userId"
## Joining, by = "userId"
## Joining, by = "userId"
## Joining, by = "userId"
```

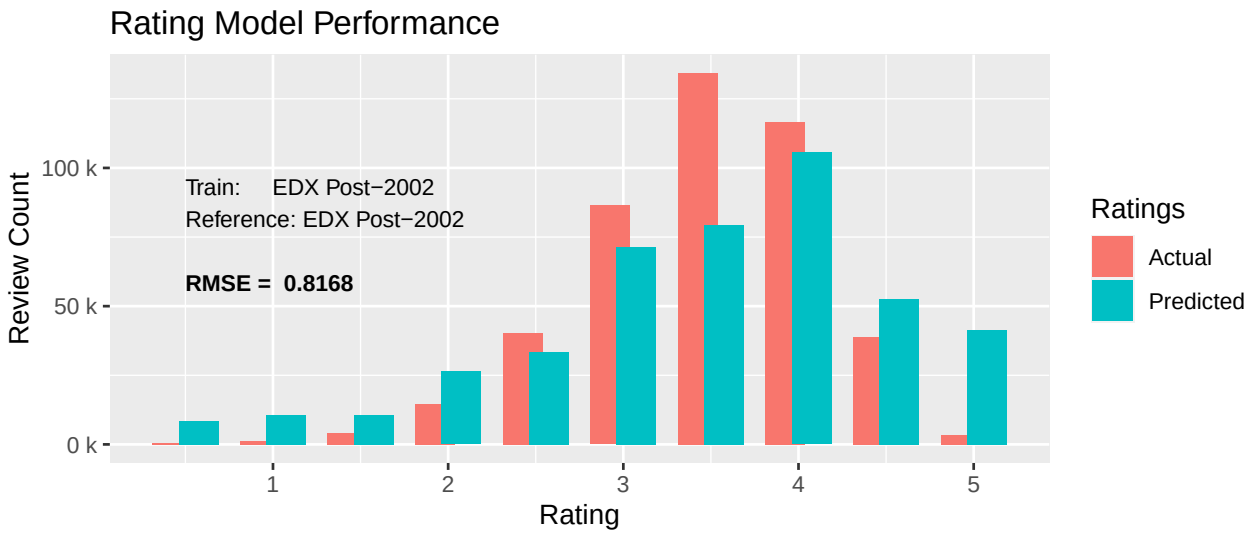
Looking at the graphs, the predictions for whole-point ratings were relatively good. The half-point ratings (i.e. 2,5, 3.5, 4.5) were under-estimated. This points to a need to modify the data set to remove these holes from the training input as pointed out in the analysis section of this document. To understand this better, the model was trained with data more recent than 2002, which contains half-point ratings at expected ratios. Prior to 2002, the data does not contain half-pints. The results show in the 3rd graph below that the RMSE fell from 0.862 to 0.817, a significant improvement. This indicates that the training data needs half-point data, real or simulated. This requires further study.

The model also over-estimated the ratings < 2 and badly for ratings of 5. The model may lower these predictions because it depends on averages in for predicting the ratings which distort the distribution. Alternative treatment might help here. Again, topic for further study

The issues above should have the highest priority. Once solved, additional enhancements around file age or popularity could further enhance accuracy.







Conclusion

This capstone project provided good real-world experience managing, designing and testing data science projects. There are many components that must come together to execute a complicated data analysis program including version control, model architectures, performance issues, and final packaging. This was very good experience.

We only scratched the surface for the movie rating prediction model. There are many moving parts and interrelationships. The project gave me first-hand experience grooming and analyzing a large data set. I know that the problems will increase as data set grow in size. I look forward to these challenges.